

Введение в байесовскую статистику

Опубликовано 15 февраля 2025 г.

Содержание

1 Независимость событий	3
1.1 Независимые события	3
1.2 Зависимые события	5
1.3 Пример с медицинскими тестами	9
1.4 Формула полной вероятности	15
2 Формула Байеса	18
2.1 Доказательство	19
2.2 Терминология и нотация	19
2.3 Смысл формулы	21
2.4 Коэффициент Байеса	28
3 Примеры	31
3.1 Колоды карт	31
3.2 Игровые кости	33
4 Оценка распределения	39
4.1 Пять мешков	40
4.2 Правдоподобие и вероятность	47
4.3 100 мешков	48
4.4 Бесконечное количество гипотез	57
5 Биномиальное и бета-распределение	63
5.1 Бета-распределение	63
5.2 Бета- и биномиальное сопряженные распределения	65
5.3 Бета-биномиальное распределение	69
5.4 Мода априорного и апостериорного распределений	73
5.5 Прогнозное априорное распределение	75
5.6 Прогнозное апостериорное распределение	77
6 Сопряженные нормальные распределения	78
6.1 Правдоподобие	78
6.2 Априорное распределение	79
6.3 Апостериорное распределение	79
6.4 Апостериорные параметры	82
6.5 Прогнозное априорное распределение	83
6.6 Прогнозное апостериорное распределение	83
7 Процесс Пуассона	85
7.1 Распределение Пуассона	85
7.2 Процесс (поток) Пуассона	94
7.3 Экспоненциальное распределение	94
7.4 Время ожидания k-ого события	107
7.4.1 Распределение Эрланга	108
7.4.2 Гамма-распределение	109
7.5 Сопряженность распределений	114
7.6 Прогнозные распределения	117

Статистика делится на описательную статистику и статистический вывод.

Статистика вывода (statistical inference) стремится сделать обоснованное предположение о параметрах генеральной совокупности на основе ограниченного набора данных (выборки).

Байесовская статистика (Bayesian inference), как один из подходов статистики вывода, использует формулу или теорему Томаса Байеса для того, чтобы уточнить изначальное представление о неизвестном параметре.

1 Независимость событий

1.1 Независимые события

Совместная вероятность. Найдем вероятность одновременного наступления двух независимых событий A и B .

Также говорят про вероятность **пересечения** (intersection) событий или **совместную вероятность** (joint probability) и обозначают $P(A \cap B)$ или $P(A, B)$.

Предположим, что события A и B означают выпадение решки (Head, H) при двукратном подбрасывании симметричной монеты (монета одна и та же, события независимы, вероятность решки или орла равна 0,5).

Интуитивное объяснение. Интуитивно (рисунок 1) речь идет об отношении количества благоприятствующих событию $A \cap B$ исходов (такой исход один) к количеству всех возможных исходов (их четыре).

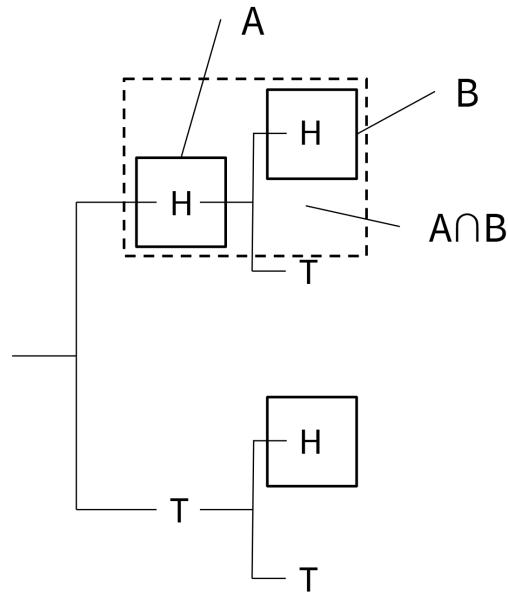


Рис. 1: Совместная вероятность независимых событий

Теорема умножения вероятностей. С другой стороны, если событию A благоприятствуют один из двух исходов и событию B также благоприятствуют один из двух исходов,

то событиям A и B , то есть $A \cap B$, благоприятствуют один из четырех исходов.

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

С точки зрения вероятности этих событий можно записать, что

$$P(H_1 \cap H_2) = P(H_1) \cdot P(H_2) = 0,5 \cdot 0,5 = 0,25$$

Таким образом, общая формула для совместной вероятности независимых событий выглядит так

$$P(A \cap B) = P(A) \cdot P(B)$$

Эту формулу называют **теоремой умножения вероятностей** (probability multiplication rule).

Биномиальное распределение. Случайную величину такого испытания можно также описать с помощью биномиального распределения, в котором $n = 2$, $p = q = 0,5$, а событие $A \cap B$ означает, что случайная величина X дважды примет значение решки, то есть $X = 2$.

Тогда по формуле бинома Ньютона

$$\begin{aligned} & \binom{2}{0} p^{2-0} q^0 + \binom{2}{1} p^{2-1} q^1 + \binom{2}{2} p^{2-2} q^2 = \\ & \binom{2}{0} 0,5^2 \cdot 0,5^0 + \binom{2}{1} 0,5^1 \cdot 0,5^1 + \binom{2}{2} 0,5^0 \cdot 0,5^2 = \\ & 1 \cdot 0,25 + 2 \cdot 0,25 + 1 \cdot 0,25 = 0,25 + 0,5 + 0,25 = 1 \end{aligned}$$

Первый член бинома как раз соответствует вероятности выпадения двух решек. Приведем график и таблицу распределения (рисунок 2).

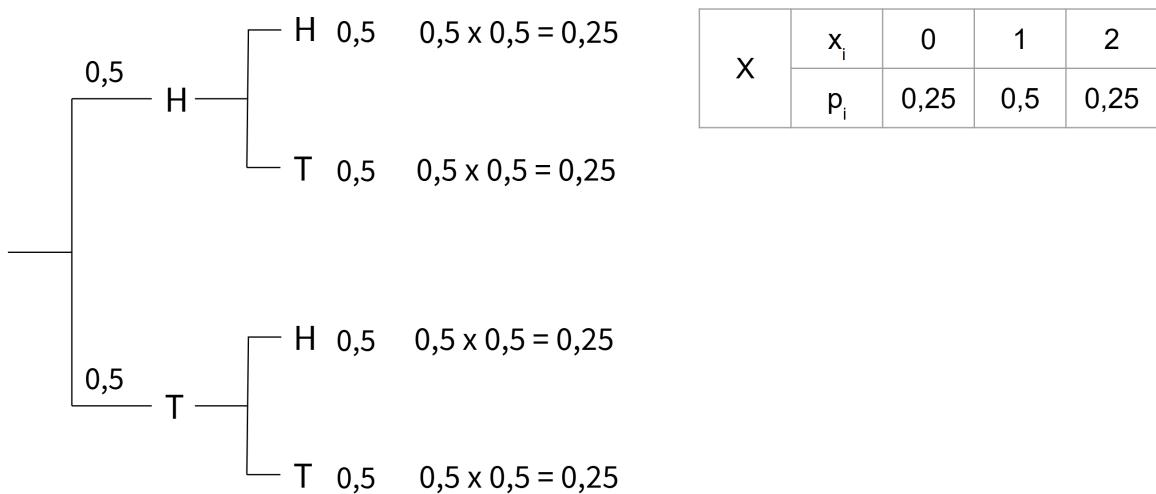


Рис. 2: Выпадение двух решек при одинаковой вероятности

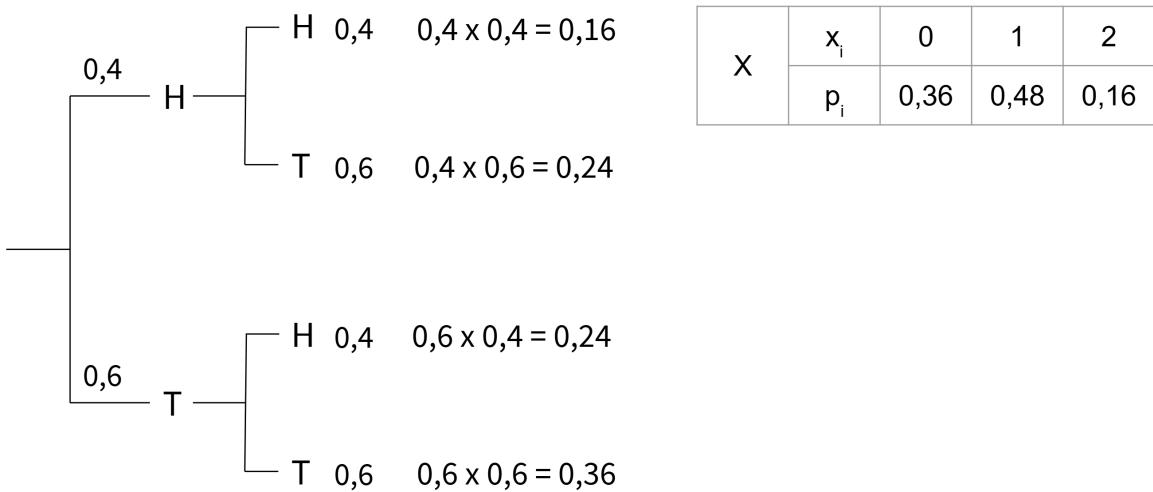


Рис. 3: Выпадение двух решек при разной вероятности

Разные вероятности. Теперь дополним картину разной вероятностью выпадения решки и орла (рисунок 3). Например, 0,4 и 0,6. По формуле совместной вероятности найдем, что

$$P(H_1 \cap H_2) = P(H_1) \cdot P(H_2) = 0,4 \cdot 0,6 = 0,16$$

По формуле бинома

$$\binom{2}{0} 0,4^2 \cdot 0,6^0 + \binom{2}{1} 0,4^1 \cdot 0,6^1 + \binom{2}{2} 0,4^0 \cdot 0,6^2 = \\ 1 \cdot 0,16 + 2 \cdot 0,4 \cdot 0,6 + 1 \cdot 0,36 = 0,16 + 0,48 + 0,36 = 1$$

Изменим условия испытаний.

1.2 Зависимые события

Однаковые вероятности. Рассмотрим подбрасывание двух разных монет с вероятностью выпадения решки на каждой из них, равной 0,5 (рисунок 4).

На этот раз предположим, что эти события зависимы. Другими словами, вообразим, что исход подбрасывания второй монеты зависит от исхода подбрасывания первой.

Примерами зависимых событий в реальной жизни будут вероятность получения штрафа в зависимости от стиля езды или получения хорошей оценки на экзамене в зависимости от уровня подготовки.

Для наглядности представим, что мы сделали двадцать серий из двух бросков каждой монеты и записали результаты в **таблицу сопряженности** (contingency table) как в виде абсолютных значений, так и в виде вероятностей (рисунок 5).

В таблице хорошо видно, что для каждой из монет M1 и M2 мы получили по 10 решек и орлов. Это так называемые маргинальные частоты или вероятности (marginal probabilities), поскольку они находятся «на полях» (от англ. margin) таблицы, а не в ее центре.

Поясним, что

- маргинальные частоты по строке показывают результат подбрасывания первой монеты M1;
- маргинальные частоты по столбцу показывают результат подбрасывания M2.

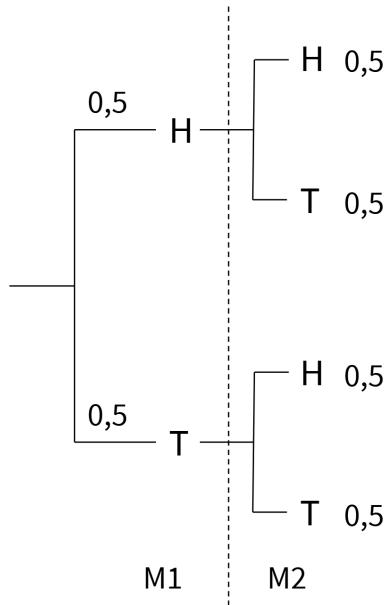


Рис. 4: Совместная вероятность зависимых событий

		M2		
		H	T	
M1	H	5 5/20 = 0,25	5 5/20 = 0,25	10 10/20 = 0,5
	T	5 5/20 = 0,25	5 5/20 = 0,25	10 10/20 = 0,5
		10 10/20 = 0,5	10 10/20 = 0,5	20 20/20 = 1

абсолютная частота

маргинальные частоты по строкам;
подбрасывание M1

маргинальные частоты по столбцам;
подбрасывание M2

относительная частота

Рис. 5: Таблица сопряженности двух зависимых событий с одинаковой вероятностью

Так как выпавшие на первой монете M1 10 решек также распределяются поровну при подбрасывании второй монеты, то совместная вероятность $P(A \cap B)$ будет также равна

$$P(A \cap B) = \frac{10}{20} \cdot \frac{5}{10} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0,25,$$

что соответствует теореме умножения вероятностей $P(A \cap B) = 0,5 \cdot 0,5 = 0,25$.

Условная вероятность. Однако можно сказать, что мы считали выпадения решки на второй монете (H_2) при условии выпадения решки на первой (H_1). Такая вероятность

называется **условной** (conditional) и записывается как $P(H_2 | H_1)$. То есть

$$P(A \cap B) = P(H_1) \cdot P(H_2 | H_1)$$

Отсюда несложно вывести общую формулу

$$P(A \cap B) = P(A) \cdot P(B | A)$$

Геометрически (рисунок 6), из общего числа бросков первой монеты мы взяли только те, при которых выпала решка (их было 10), и внутри этого вероятностного пространства на второй монете решка выпала также в половине случаев (или пять раз).

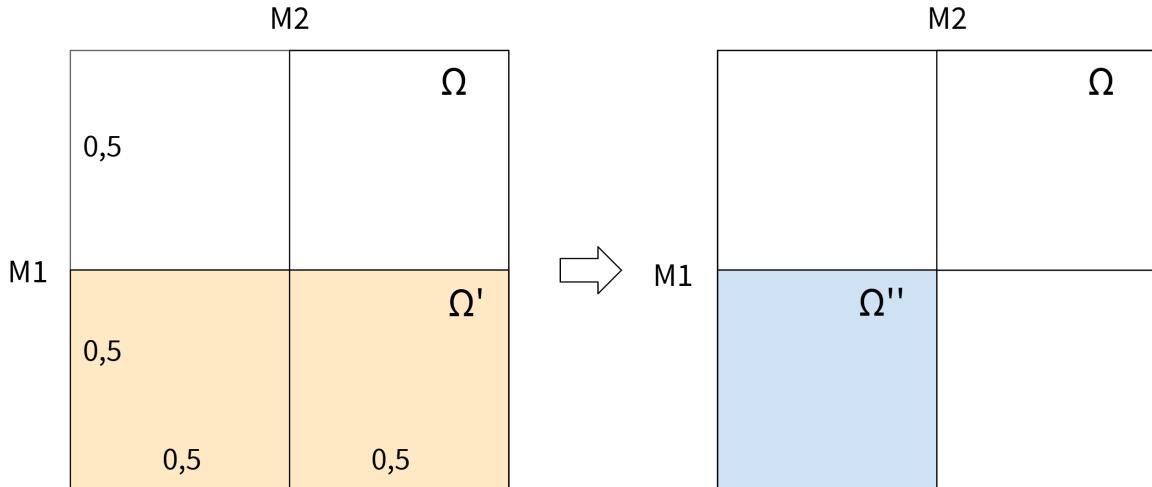


Рис. 6: Два зависимых события с одинаковой вероятностью

Можно представить, что вероятности события A «ограничивают» пространство исходов Ω (весь квадрат) и вероятности события B «довольствуются» тем, что осталось в каждой из половинок. Мы как бы сокращаем варианты выбора для второй монеты подбрасыванием первой монеты (Ω'), а затем берем «половину от половины» (Ω'').

Примечание. Заметим, что если события независимы, то вероятность события B при условии A равна просто вероятности события B , $P(B | A) = P(B)$.

Теперь, так как у нас две монеты, поэкспериментируем с разными вероятностями выпадения решки или орла на каждой из них.

Разные вероятности. Зададим новые условия (рисунок 7):

- пусть вначале мы подбрасываем монету, которая выпадает решкой с вероятностью 0,4, а орлом соответственно с вероятностью 0,6;
- после этого мы подбрасываем обычную симметричную монету с вероятностями 0,5.

Поставим задачу найти вероятность того, что на второй монете выпала решка (H_2), при условии, что на первой монете выпал орел (T_1). Другими словами, найти вероятность $P(T_1 \cap H_2)$ или, если использовать запись условной вероятности, $P(T_1) \cdot P(H_2 | T_1)$.

Посмотрим (рисунок 8), как это отразилось на пространстве исходов.

Несложно рассчитать, что

$$P(T_1 \cap H_2) = P(T_1) \cdot P(H_2 | T_1) = 0,6 \cdot 0,5 = 0,30$$

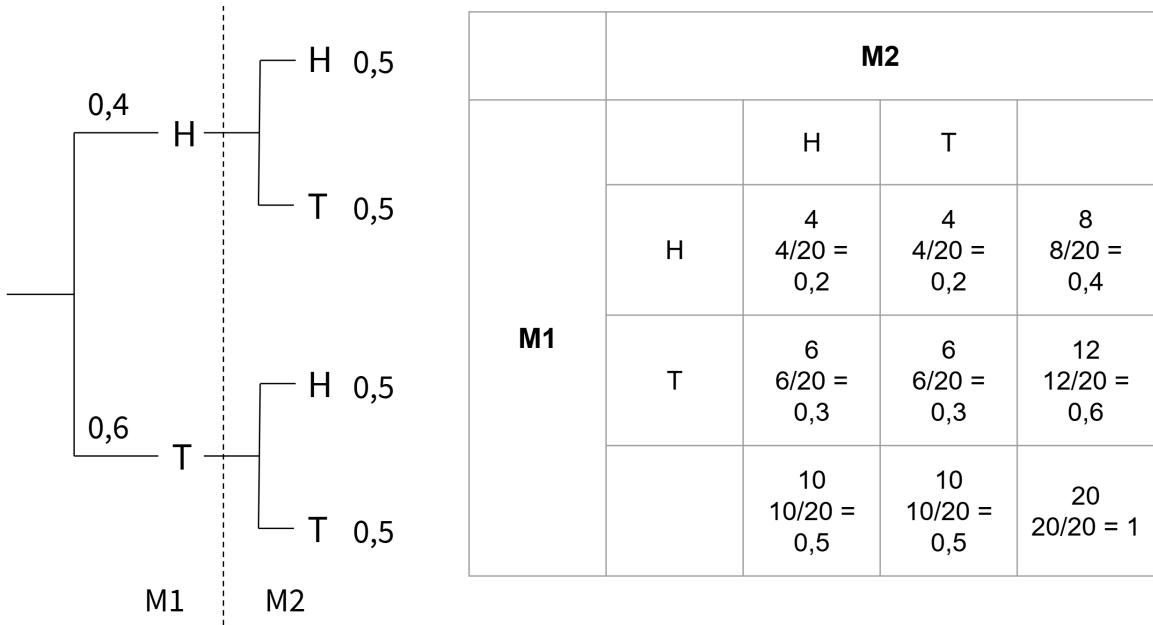


Рис. 7: Таблица сопряженности двух зависимых событий с разной вероятностью

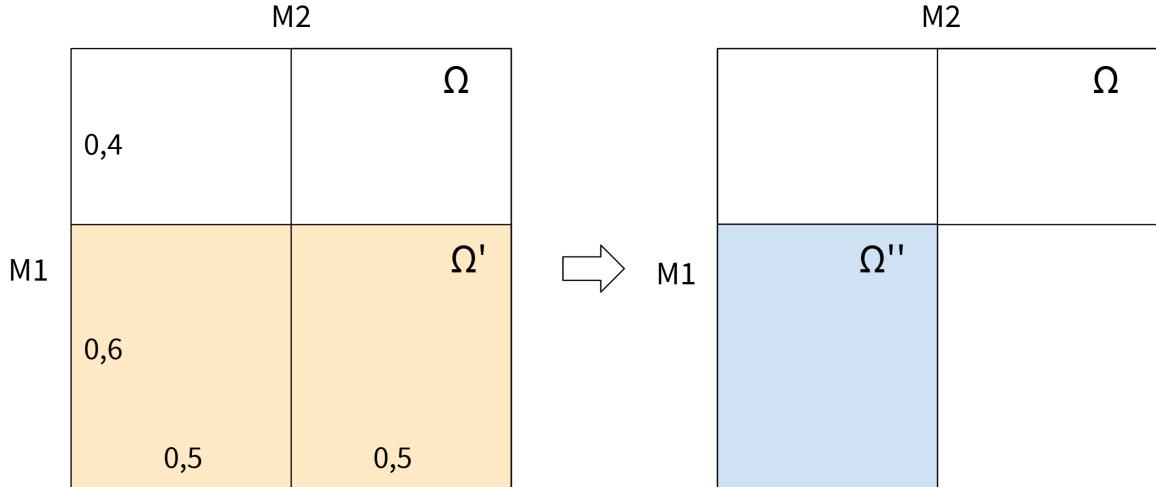


Рис. 8: Два зависимых события с разной вероятностью

Как мы видим, вероятность уже не равна 0,25, на совместную вероятность двух событий повлияло то, что первая монета несимметрична. Геометрически после выпадения второй монеты мы нашли половину от большей площади пространства исходов Ω .

Пойдем дальше в наших экспериментах и предположим, что

- на первой монете решка выпадает с вероятностью 0,2, а на второй монете вероятности решки и орла равны 0,8 и 0,2 соответственно; а вот
- если выпадает орел на первой монете (с вероятностью 0,8), то наоборот, вероятности решки и орла на второй равны 0,2 и 0,8.

Приведем диаграмму (рисунок 9). Найдем вероятность выпадения двух решек $P(H_1, H_2)$.

$$P(H_1 \cap H_2) = P(H_1) \cdot P(H_2 | H_1) = 0,2 \cdot 0,8 = 0,16$$

Примечательно, что хотя вероятность выпадения решки на второй монете очень велика

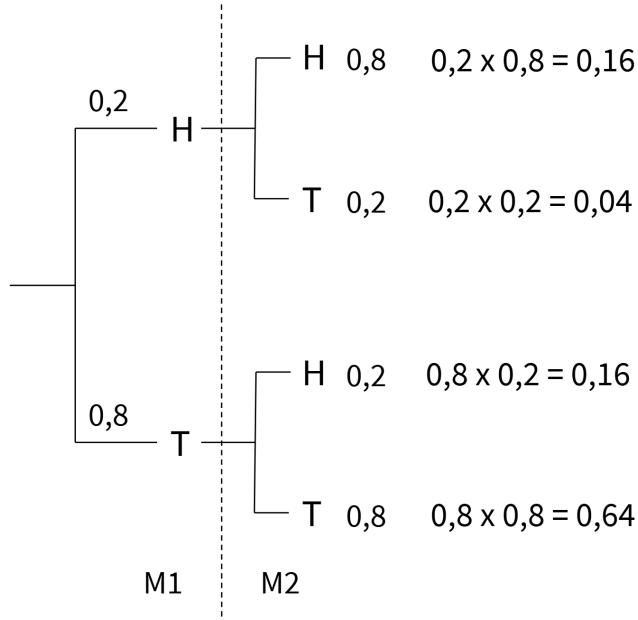


Рис. 9: Новые вероятности зависимых событий. Диаграмма

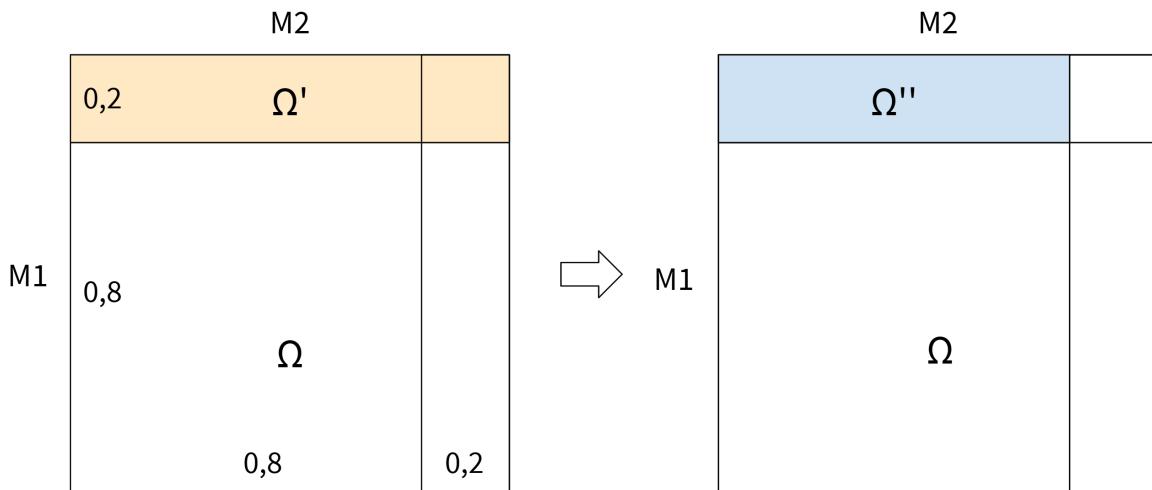


Рис. 10: Новые вероятности зависимых событий. Пространство исходов

(0,8), такое событие как подбрасывание второй монеты после решки на первой в принципе маловероятно (0,2).

Геометрически (рисунок 10), небольшая вероятность выпадения решки при первом подбрасывании существенно ограничивает вероятность выпадения решки при втором и здесь, вне зависимости от величины этой вероятности, пусть даже 0,99, совместная вероятность двух событий будет существенно ниже.

1.3 Пример с медицинскими тестами

Это интересное свойство совместной вероятности зависимых событий часто иллюстрируют с помощью медицинских тестов. Рассмотрим некоторое заболевание (какое именно значения не имеет), для диагностики которого разработан медицинский тест.

Матрица ошибок. Из вводного курса ML (занятие по классификации) мы знаем, что классификатор может делать:

- истинно положительные (TP);
- истинно отрицательные (TN);
- ложно положительные (FN); и наконец
- ложно отрицательные (FP) прогнозы.

Эти результаты (рисунок 11) мы объединили в **матрицу ошибок** (confusion matrix).

	Тест: отрицательный (predicted negative)	Тест: положительный (predicted positive)
Факт: отрицательный (actual negative)	Истинно отрицательный (true negative, TN)	Ложноположительный (false positive, FP)
Факт: положительный (actual positive)	Ложноотрицательный (false negative, FN)	Истинно положительный (true positive, TP)

Рис. 11: Матрица ошибок классификатора

Качество медицинских тестов принято оценивать по двум критериям, которые можно рассчитать с помощью показателей этой матрицы.

Чувствительность теста. Во-первых, **чувствительность** (sensitivity) теста или доля истинно положительных результатов (true positive rate, TPR) определяется способностью выдавать положительный прогноз в случае, когда человек действительно болен. С точки зрения матрицы ошибок речь идет об отношении TP к $TP + FN$.

$$TPR = \frac{TP}{TP + FN}$$

Другими словами, мы сравниваем тех, кто действительно болен и показал истинно положительный тест (TP) с суммой этих людей, а также тех, для кого тест показал ложноотрицательный результат ($TP + FN$), поскольку они тоже больны.

Почему это важно? Если этот показатель будет низким, то мы не сможем выявить многих действительно заболевших и не начнем лечение.

Заметим, что если рассматривать матрицу ошибок как таблицу сопряженности (рисунок 12), то фактически чувствительность теста это отношение тех, у кого положительный тест к маргинальной вероятности тех, кто действительно болен.

Таким образом, с точки зрения теории вероятностей чувствительность — это совместная вероятность быть больным (первая монета из примера выше) и одновременно показать положительный результат на teste (вторая монета).

$$P(\text{чувствительность}) = P(\text{болен} \cap +) = P(\text{болен}) \cdot P(+ | \text{болен})$$

Посмотрим на диаграмму (рисунок 13).

	Тест: —	Тест: +
Факт: не болен	Истинно отрицательный (TN)	Ложноположительный (FP)
Факт: болен	Ложноотрицательный (FN)	Истинно положительный (TP)

маргинальная вероятность
быть больным

Рис. 12: Доля истинно положительных тестов (TPR)

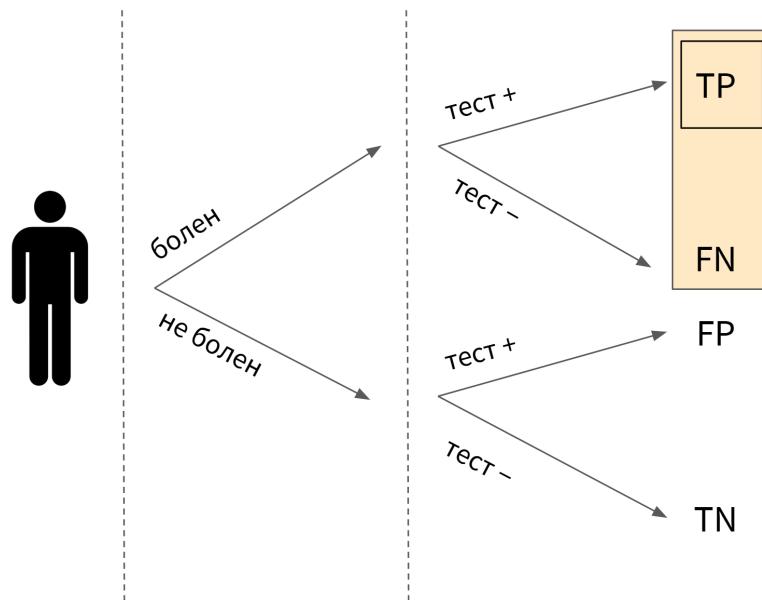


Рис. 13: TPR на диаграмме

Специфичность теста. Во-вторых, **специфичность** (specificity) теста или доля истинно отрицательных результатов (true negative rate, TNR) — это способность теста показывать отрицательный результат, в случае когда человек действительно здоров (рисунки 14 и 15).

$$TNR = \frac{TN}{TN + FP}$$

Говоря простыми словами, это способность теста не поднимать лишнюю панику. Если у теста низкая специфичность, положительные результаты будут часто получать как больные, так и здоровые люди.

Тогда,

$$P(\text{специфичность}) = P(\text{не болен} \cap -) = P(\text{не болен}) \cdot P(- | \text{не болен})$$

Распространенность заболевания. Наконец, **распространенность** (prevalence) заболевания определяется, говоря упрощенно, как доля заболевших к численности насе-

	Тест: —	Тест: +
Факт: не болен	Истинно отрицательный (TN)	Ложноположительный (FP)
Факт: болен	Ложноотрицательный (FN)	Истинно положительный (TP)

маргинальная вероятность быть не больным

Рис. 14: Доля истинно отрицательных результатов (TNR)

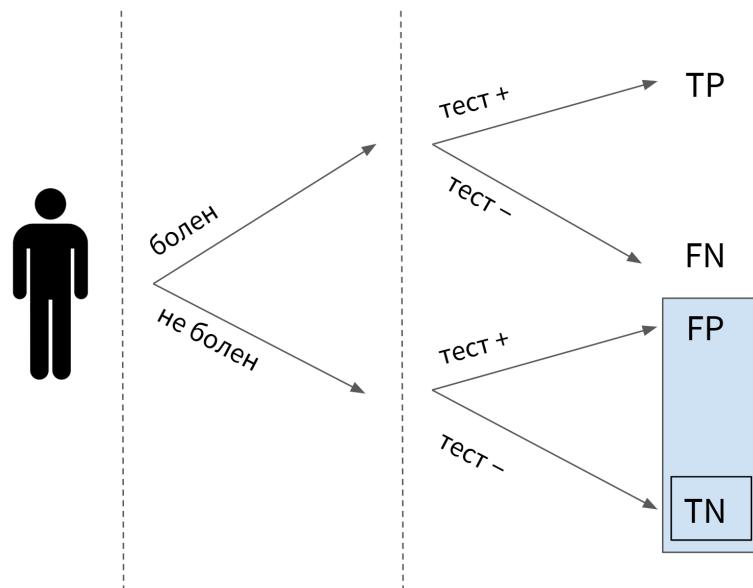


Рис. 15: TNR на диаграмме

ния.

Почему это важно в наших расчетах? Распространенность будет играть роль первой монеты. Это вероятность того, что человек болен до сдачи теста. Вторая монета — это тест, вернее его чувствительность и специфичность (рисунок 16).

Вероятность оказаться больным. Рассмотрим конкретный пример. Пусть

- распространенность заболевания ограничена пятью процентами населения;
- чувствительность теста составляет 90 процентов; а
- специфичность — 80 процентов.

Найдем вероятность того, что случайный человек, получивший положительный результат теста окажется действительно болен, то есть $P(\text{болен} | +)$. Вначале заполним диаграмму полученными вероятностями (рисунок 17).

Логично, что для того чтобы найти вероятность быть больным при положительном teste,

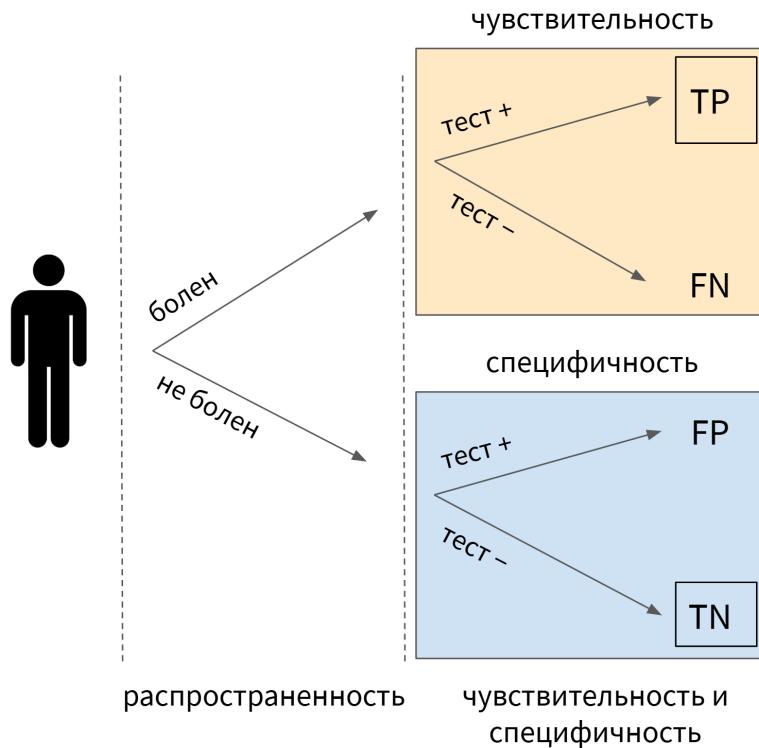


Рис. 16: Распространенность, чувствительность и специфичность

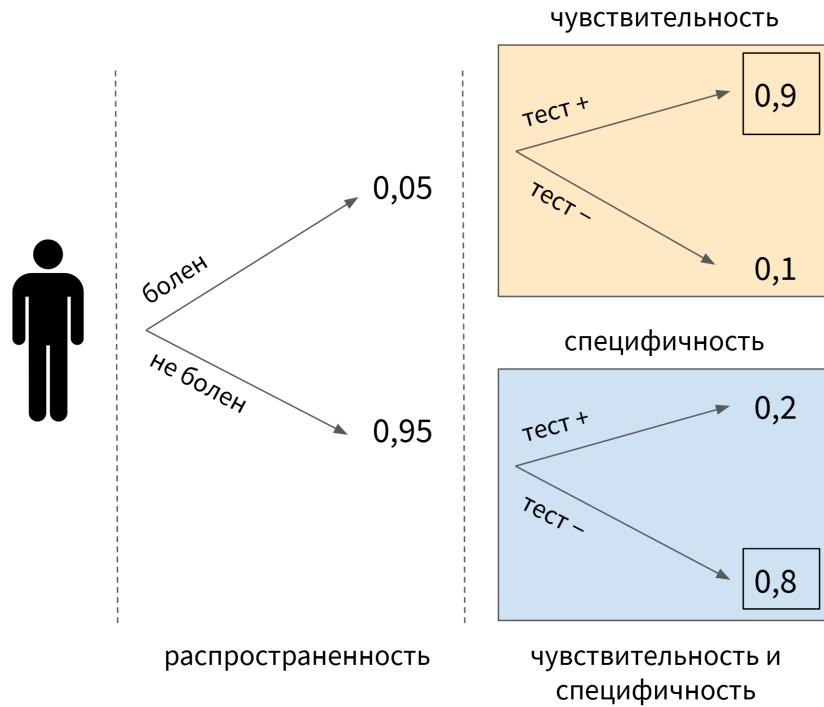


Рис. 17: Вероятность оказаться больным

нужно найти отношение совместной вероятности быть больным и получить положительный тест к вероятности получить положительный тест (ведь положительный тест получают не только больные, но и здоровые люди). Другими словами,

$$P(\text{болен} | +) = \frac{P(\text{болен} \cap +)}{P(+)}$$

Начнем с числителя. По сути, речь идет о чувствительности теста.

$$P(\text{болен} \cap +) = P(\text{болен}) \cdot P(+ | \text{болен}) = 0,05 \cdot 0,90$$

Теперь найдем знаменатель или всех тех, кто получил положительный тест. Механически можно проследить ветви диаграммы (рисунок 18)

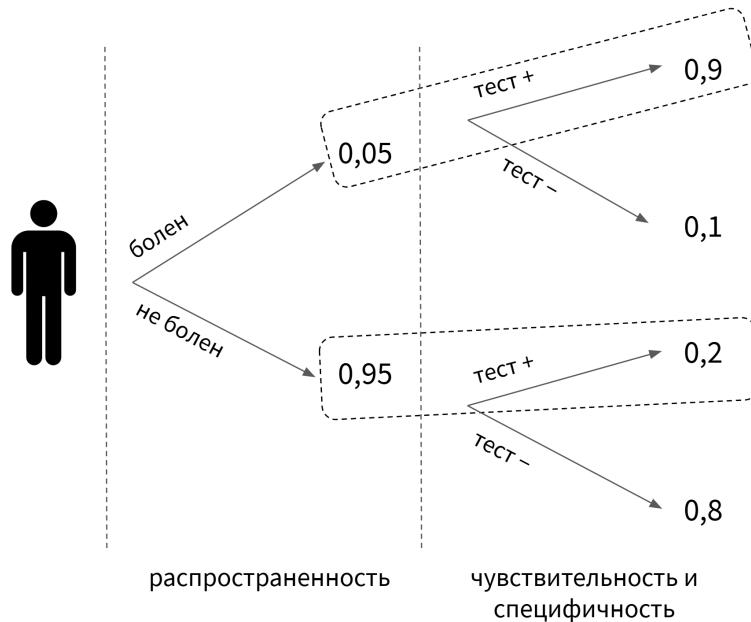


Рис. 18: Вероятность оказаться больным. Диаграмма

и перемножить соответствующие вероятности

$$P(+) = 0,05 \cdot 0,90 + 0,95 \cdot 0,20.$$

С точки зрения смысла речь идет о тех,

- кто получил положительный тест и действительно болен, то есть $P(\text{болен} \cap +)$ или $P(\text{болен}) \cdot P(+ | \text{болен})$; а также о тех,
- кто получил положительный тест и при этом не болен, то есть $P(\text{не болен} \cap +)$ или $P(\text{не болен}) \cdot P(+ | \text{не болен})$.

Соберем все это в общую формулу

$$P(\text{болен} | +) = \frac{P(\text{болен} \cap +)}{P(+)}$$

$$P(\text{болен} | +) = \frac{P(\text{болен} \cap +)}{P(\text{болен} \cap +) + P(\text{не болен} \cap +)}$$

$$P(\text{болен} | +) = \frac{P(\text{болен}) \cdot P(+ | \text{болен})}{P(\text{болен}) \cdot P(+ | \text{болен}) + P(\text{не болен}) \cdot P(+ | \text{не болен})}$$

Для наглядности свяжем термины с вероятностями (рисунки 19 и 20).

Остается вычислить вероятность

$$P(\text{болен} | +) = \frac{0,05 \cdot 0,90}{0,05 \cdot 0,90 + 0,95 \cdot 0,20} \approx 0,19$$

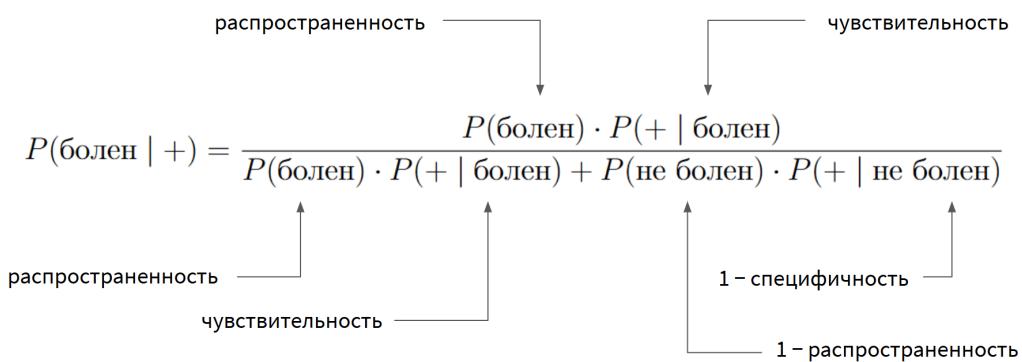


Рис. 19: Показатели медицинских тестов и вероятности

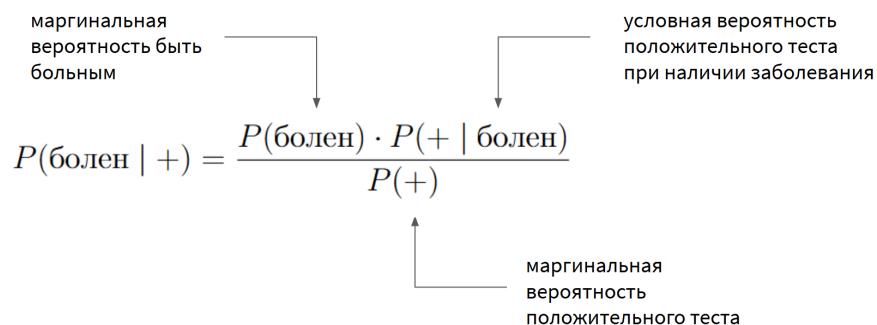


Рис. 20: Показатели медицинских тестов и вероятности. Продолжение

Мы получили неожиданный результат. При чувствительности теста в 90 процентов, его положительный результат означает, что человек действительно болен лишь с вероятностью около 19 процентов.

Обратимся к квадрату пространства исходов. Вначале найдем числитель (рисунок 21).

$$P(\text{болен}) \cdot P(+) \mid \text{болен}$$

Теперь знаменатель (рисунок 22).

$$P(\text{болен}) \cdot P(+ | \text{болен}) + P(\text{не болен}) \cdot P(+ | \text{не болен})$$

Посмотрим на общую геометрию формулы (рисунок 23).

1.4 Формула полной вероятности

Формула в общем виде. Еще раз рассмотрим знаменатель в формуле выше

$$P(+) = P(\text{болен} \cap +) + P(\text{не болен} \cap +)$$

$$P(+) = P(\text{болен}) \cdot P(+ | \text{болен}) + P(\text{не болен}) \cdot P(+ | \text{не болен})$$

Его принято называть **формулой полной вероятности** (law of total probability). Запишем эту формулу в общем виде. Назовем (рисунок 24):

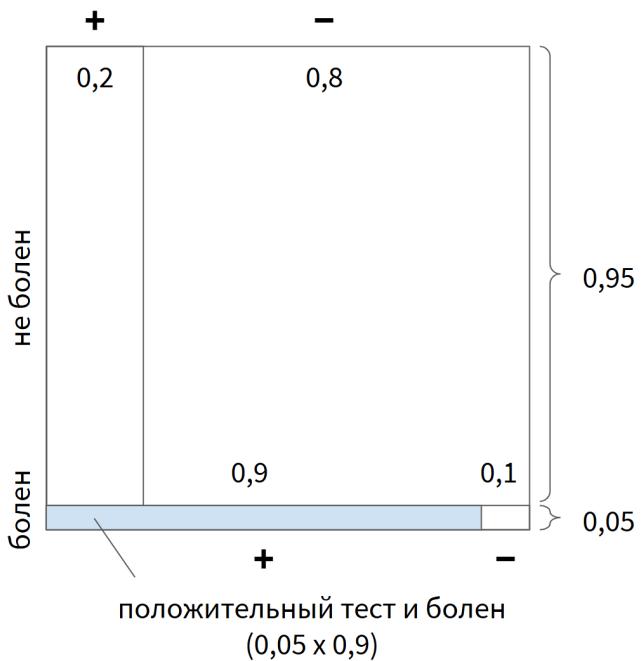


Рис. 21: Числитель вероятности быть больным при положительном тесте

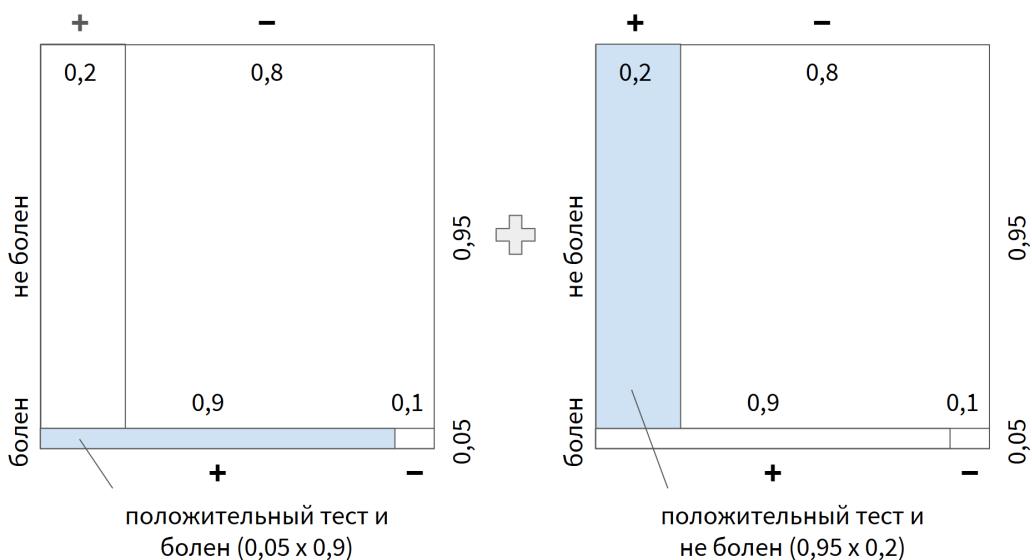


Рис. 22: Знаменатель вероятности быть больным при положительном тесте

- возможность заболеть событием A ,
 - тогда событием A_1 будет то, что человек болен; а
 - событием A_2 — то, что человек здоров; при этом
- положительный тест обозначим событием B .

Примечание: событие В включает только случаи положительного теста, при этом событие А включает как случаи заболевания (A_1), так и его отсутствия (A_2).

В этом случае

$$P(B) = \sum_n A_n \cap B$$

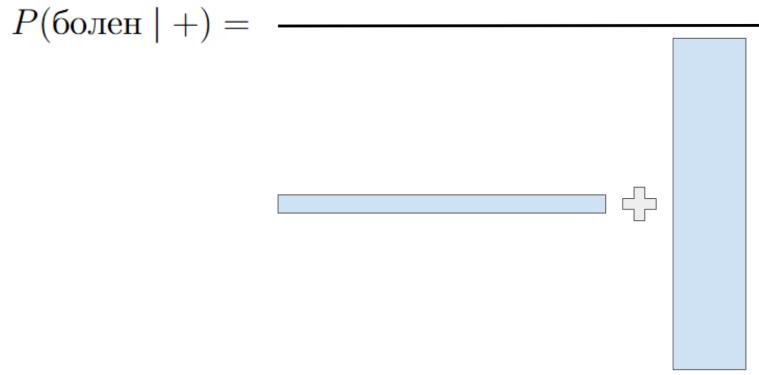


Рис. 23: Вероятность быть больным при положительном тесте

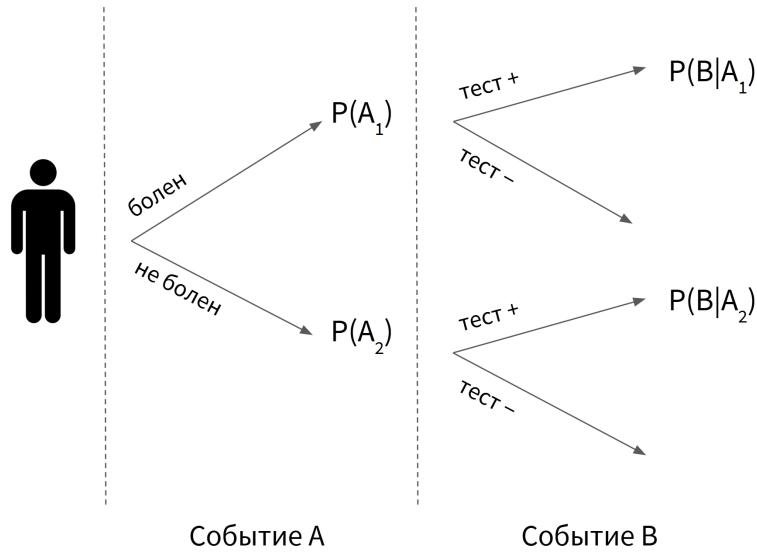


Рис. 24: Компоненты формулы полной вероятности

Или, что то же самое

$$P(B) = P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)$$

$$P(B) = \sum_n P(A_n) \cdot P(B | A_n)$$

Вопрос нотации. Часто недопонимание формул возникает из-за неуточненной нотации. В частности, какое событие считать событием A , а какое событием B . Если событием B считать факт заболевания, а событием A — результат теста, то формула полной вероятности будет выглядеть так

$$P(A) = \sum_n A \cap B_n$$

$$P(A) = \sum_n P(A | B_n) \cdot P(B_n)$$

Кроме того, обратите внимание, что во второй формуле мы поменяли множители местами. От этого результат никак не изменится.

В примере с медицинскими тестами нас интересовала вероятность того, что человек болен, при условии положительного теста $P(\text{болен}) \cdot P(+ | \text{болен})$. Знаменатель нормализовал это значение, включив те случаи, когда как больные, так и здоровые люди получали положительный результат $P(+)$.

Если бы мы хотели сравнить вероятность, что человек болен при условии положительного теста, с вероятностью того, что он не болен при условии также положительного теста, то формулы в развернутом виде выглядели бы так

$$P(\text{болен} | +) = \frac{P(\text{болен}) \cdot P(+ | \text{болен})}{P(\text{болен}) \cdot P(+ | \text{болен}) + P(\text{не болен}) \cdot P(+ | \text{не болен})}$$

$$P(\text{не болен} | +) = \frac{P(\text{не болен}) \cdot P(+ | \text{не болен})}{P(\text{болен}) \cdot P(+ | \text{болен}) + P(\text{не болен}) \cdot P(+ | \text{не болен})}$$

В обоих случаях знаменатель был бы одинаковый, $P(+)$, или как еще говорят, играл роль нормализующей константы, $const$, а значит он никак не повлиял бы на сравнение.

Вначале вычислим вероятности с учетом знаменателя

$$P(\text{болен} | +) = \frac{0,05 \cdot 0,90}{0,05 \cdot 0,90 + 0,95 \cdot 0,20} \approx 0,19$$

$$P(\text{не болен} | +) = \frac{0,95 \cdot 0,20}{0,05 \cdot 0,90 + 0,95 \cdot 0,20} \approx 0,81$$

Теперь сравним только числители, которые будут пропорциональны этим вероятностям

$$\begin{aligned} P(\text{болен} | +) &\propto 0,05 \cdot 0,90 = 0,045 \\ P(\text{не болен} | +) &\propto 0,95 \cdot 0,20 = 0,19 \end{aligned}$$

В некоторых ситуациях, как мы увидим позднее, от вычисления знаменателя (т.е. формулы полной вероятности) бывает удобно отказаться.

2 Формула Байеса

Выше мы научились находить условную вероятность зависимых событий. Например, вероятности заболевания при условии положительного теста $P(\text{болен} | +)$. По сути получившаяся формула и есть формула Байеса.

Тем не менее для того чтобы двигаться дальше будет полезно:

- формально доказать эту теорему;
- ввести необходимую терминологию и нотацию; а также
- познакомиться с основной идеей байесовской статистики, которая заключается в возможности многократного применения формулы Байеса для получения более достоверной условной вероятности.

2.1 Доказательство

Выше мы показали, что совместная вероятность зависимых событий находится по формуле

$$P(A \cap B) = P(A) \cdot P(B | A)$$

$$P(B \cap A) = P(B) \cdot P(A | B)$$

При этом совместная вероятность события A и события B равна совместной вероятности события B и события A . Другими словами,

$$P(A \cap B) = P(B \cap A)$$

Формально, если представить маргинальные вероятности событий A и B в виде множеств, то приведенное выше равенство напрямую следует из свойства коммутативности пересечения двух множеств.

$$A \cap B = B \cap A$$

Для того чтобы наглядно в этом убедиться, вернемся к подбрасыванию двух разных монет. Если мы подбрасываем М1, а затем М2, то сначала находим 0,6 площади квадрата, а затем его половину (рисунок 25).

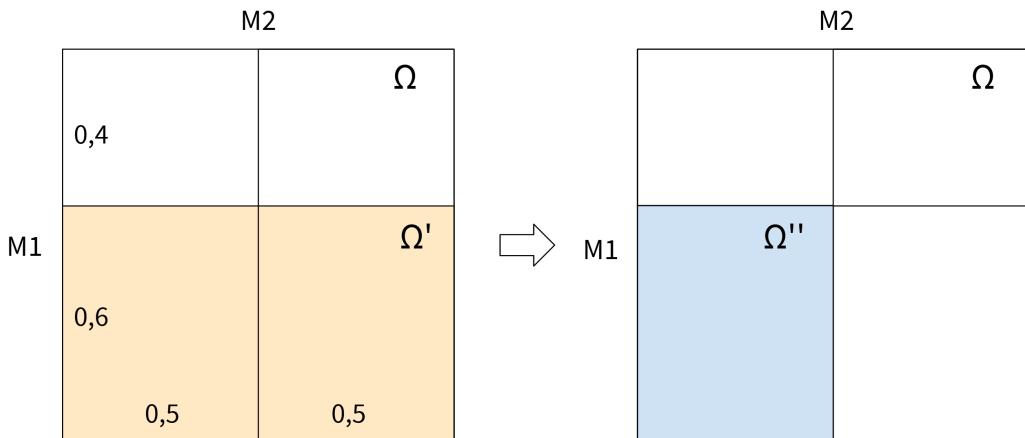


Рис. 25: Подбрасываем сначала М1, потом М2

Если сначала М2, а потом М1, то в первую очередь делим квадрат пополам, а затем находим 0,6 от первой половины (рисунок 26).

Таким образом,

$$P(A) \cdot P(B | A) = P(B) \cdot P(A | B)$$

$$P(A | B) = \frac{P(A) \cdot P(B | A)}{P(B)}$$

Это и есть **формула Байеса** (Bayes' rule, читается [beiz]) или **теорема Байеса** (Bayes' theorem).

2.2 Терминология и нотация

Основные термины. В терминологии байесовской статистики (рисунок 27) распространенность заболевания, $P(A)$, называют изначальной или **априорной вероятностью**

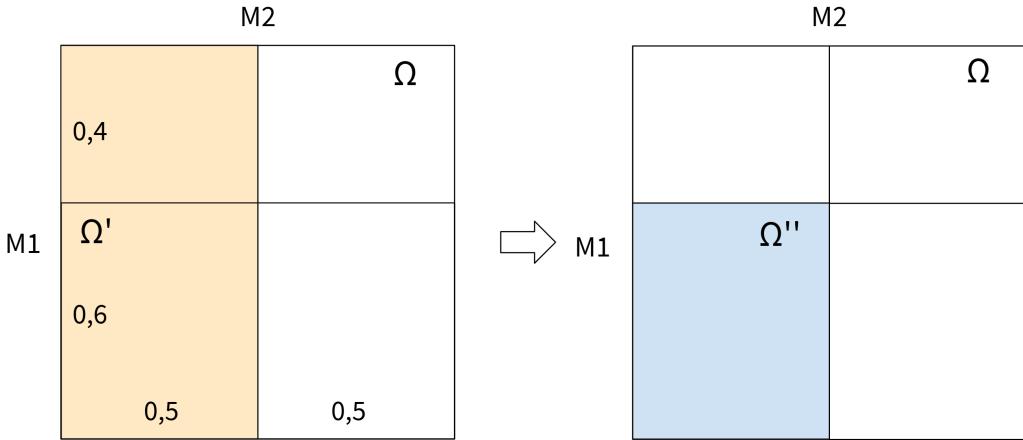


Рис. 26: Подбрасываем сначала M2, потом M1

(prior probability). Это нечто, что нам дано до того, как мы проводим какие-либо тесты или эксперименты.

Качество теста (его чувствительность и специфичность) — это то, как работает тест. В байесовской статистике это называется **правдоподобием** (likelihood).

Можно сказать, что так мы узнаем, насколько правдоподобно, что человек болен при условии положительного и отрицательного результатов, $P(B | A_1)$, и насколько правдоподобно, что он здоров, $P(B | A_2)$.

Знаменатель называют **данными** (data), которые мы получили, применяя тест к больным и здоровым людям.

Результат, который мы получаем после проведения теста, называется **апостериорной вероятностью** (posterior probability).

$$P(\text{болен} | +) = \frac{\underbrace{P(\text{болен}) \cdot P(+ | \text{болен})}_{\substack{\text{апостериорная} \\ \text{вероятность} \\ (\text{posterior})}}}{\underbrace{P(\text{болен}) \cdot P(+ | \text{болен}) + P(\text{не болен}) \cdot P(+ | \text{не болен})}_{\substack{\text{априорная} \\ \text{вероятность} \\ (\text{prior}) \quad \text{правдоподобие} \\ (\text{likelihood})}}}$$

данные
(data)

Рис. 27: Терминология формулы Байеса

В общем виде получится

$$P(A_1 | B) = \frac{P(A_1) \cdot P(B | A_1)}{P(B)}$$

$$P(A_1 | B) = \frac{P(A_1) \cdot P(B | A_1)}{P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)}$$

Hypothesis и evidence. Часто событие A также называют **гипотезой**, hypothesis, H , которая может подтвердиться, а может не подтвердиться, $\neg H$. Данные обозначают буквой

E , от англ. evidence, «полученные данные», «свидетельства» (рисунок 28).

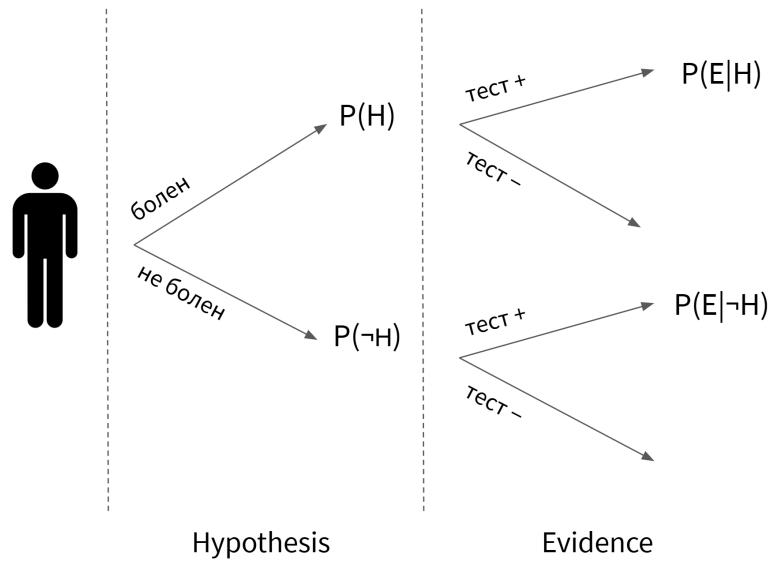


Рис. 28: Гипотеза и полученные данные

Тогда гипотезу о том, что человек болен при условии, что данные (то есть тест) об этом свидетельствуют, можно записать так

$$P(H | E) = \frac{P(H) \cdot P(E | H)}{P(E)}$$

$$P(H | E) = \frac{P(H) \cdot P(E | H)}{P(H) \cdot P(E | H) + P(\neg H) \cdot P(E | \neg H)}$$

Параметр и данные. Также, в частности, в машинном обучении событие А принято называть оцениваемым **параметром** θ , а событие В — данными или признаками X . Тогда,

$$P(\theta | X) = \frac{P(\theta) \cdot P(X | \theta)}{P(X)}$$

Далее мы будем использовать все три нотации: события A и B , гипотезу H и данные E , а также параметр θ и признаки X .

Байесовский вывод. Под байесовским выводом (bayesian inference) понимается вывод апостериорной вероятности (апостериорного распределения случайной величины, inference of a posterior distribution) на основе априорной вероятности с учетом полученных данных.

2.3 Смысл формулы

Изменение априорной вероятности и правдоподобия. Вернемся к числителям из примера с медицинскими тестами

$$P(\text{болен} | +) \propto 0,05 \cdot 0,90 = 0,045$$

$$P(\text{не болен} | +) \propto 0,95 \cdot 0,20 = 0,19$$

Что влияет на вероятность быть больным или здоровым при условии положительного теста? Опять же, знаменатель никакого влияния не окажет, поскольку в обоих случаях он одинаковый.

При этом в числителе на эту вероятность влияют две величины:

- во-первых, распространенность заболевания или вероятность быть больным (0,05 против 0,95);
- во-вторых, чувствительность и неспецифичность (то есть 1—специфичность) теста (0,9 против 0,2).

$$P(\text{болен} | +) \propto 0,05 \cdot 0,90 = 0,045$$

$$P(\text{не болен} | +) \propto 0,95 \cdot 0,20 = 0,19$$

Рис. 29: Смысл формулы Байеса

Как мы видим (рисунок 29), в данном случае «побеждает» пара вероятность быть здоровым и чувствительность против вероятности быть больным и неспецифичности.

$$0,05 \cdot 0,90 = 0,045 < 0,95 \cdot 0,20 = 0,19$$

При других показателях распространенности заболевания, чувствительности и специфичности теста результат был бы иным. Например, для более распространенного заболевания, скажем 0,5 и 0,5, вероятность быть больным при положительном тесте резко возрастает.

$$0,5 \cdot 0,90 = 0,45 > 0,5 \cdot 0,20 = 0,1$$

Рассчитаем вероятность быть больным (листинг 1) и здоровым (листинг 2) при положительном тесте с помощью Питона.

```
prior = np.array([0.05, 0.95])
likelihood = np.array([0.9, 0.2])

P_sick_positive = prior[0] * likelihood[0] / np.dot(prior, likelihood)
P_sick_positive

Output: 0.19148936170212766
```

Листинг 1: Вероятность быть больным при положительном тесте

```
P_not_sick_positive = prior[1] * likelihood[1] / np.dot(prior,
    likelihood)
P_not_sick_positive

Output: 0.8085106382978723
```

Листинг 2: Вероятность быть здоровым при положительном тесте

Обратите внимание, что в знаменателе распространенность очень удобно перемножать с чувствительностью и неспецифичностью с помощью скалярного произведения.

Геометрически (рисунок 30), изменение априорной вероятности (распространенности) и правдоподобия (качества теста) можно представить следующим образом.

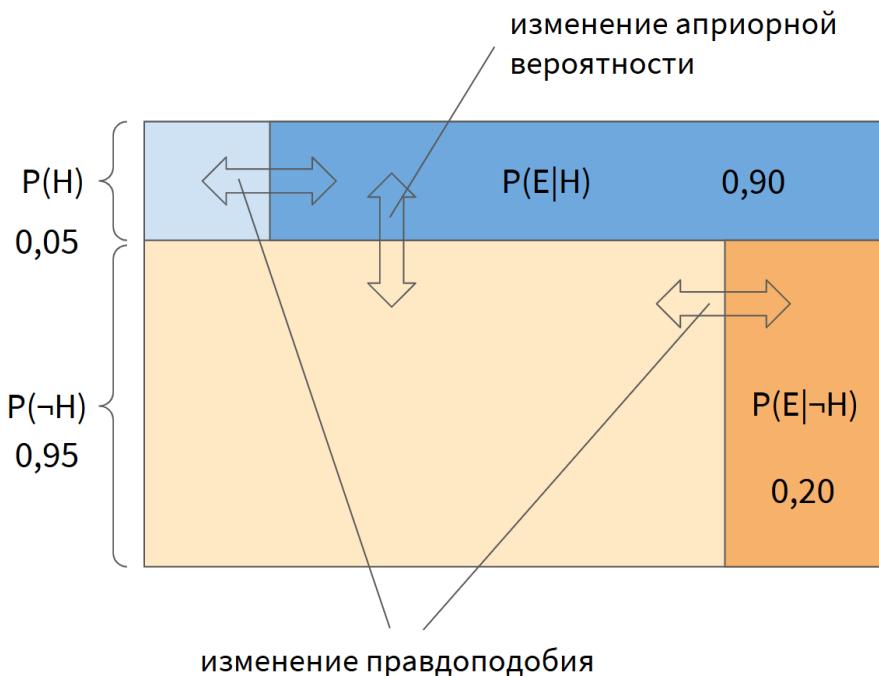


Рис. 30: Геометрия априорной вероятности и правдоподобия

Опять же, если предположить, что распространность заболевания составляет, например, лишь 0,01, то вероятность быть больным при положительном тесте должна быть ниже (листинг 3).

```
prior = np.array([0.01, 0.99])
likelihood = np.array([0.9, 0.2])

P_sick_positive = prior[0] * likelihood[0] / np.dot(prior, likelihood)
P_sick_positive

Output: 0.043478260869565216
```

Листинг 3: Меньшая распространность заболевания

Если же повысить чувствительность теста до 0,95, а его специфичность до 0,90, то вероятность быть больным при положительном результате увеличится (листинг 4).

```
prior = np.array([0.05, 0.95])
likelihood = np.array([0.95, 0.1])

P_sick_positive = prior[0] * likelihood[0] / np.dot(prior, likelihood)
P_sick_positive

Output: 0.3333333333333333
```

Листинг 4: Большая чувствительность теста

Обновление априорной вероятности. Вероятность заболевания при положительном тесте в 19 процентов на самом деле не может удовлетворять ни пациента, ни его врача. Необходимо провести повторное тестирование.

При этом здесь есть нюанс. Мы не можем не учитывать результаты первого теста (апостериорную вероятность) при повторном тестировании. Но как добавить их в новые расчеты?

Мы можем при каждом новом тесте делать апостериорную вероятность новой априорной или, как еще говорят, **обновлять априорную вероятность** (update prior).

Вновь найдем апостериорную вероятность быть больным после первого положительного теста (листинг 5).

```
prior = np.array([0.05, 0.95])
likelihood = np.array([0.9, 0.2])

posterior = prior[0] * likelihood[0] / np.dot(prior, likelihood)
posterior

Output: 0.19148936170212766
```

Листинг 5: Апостериорная вероятность после первого положительного теста

Сделаем эту апостериорную вероятность новой априорной (листинг 6).

```
new_prior = np.array([posterior, 1 - posterior])
new_prior

Output: array([0.19148936, 0.80851064])
```

Листинг 6: Новая априорная вероятность

Заметим, что $(1 - \text{апостериорная вероятность})$ — это то же самое, что $P(\text{не болен} | +)$ (листинг 7).

```
prior[1] * likelihood[1] / np.dot(prior, likelihood)

Output: 0.8085106382978723
```

Листинг 7: 1—апостериорная вероятность

Приведем схему (рисунок 31) и найдем новую апостериорную вероятность (листинг 8).

```
new_posterior = new_prior[0] * likelihood[0] / np.dot(new_prior,
    likelihood)
new_posterior

Output: 0.5159235668789809
```

Листинг 8: Новая апостериорная вероятность

Как мы видим, после двух положительных тестов вероятность быть больным составляет уже около 52 процентов.

Обновление как степень правдоподобия. Для того чтобы оценить вероятность после трех положительных тестов, можно воспользоваться циклом (листинг 9).

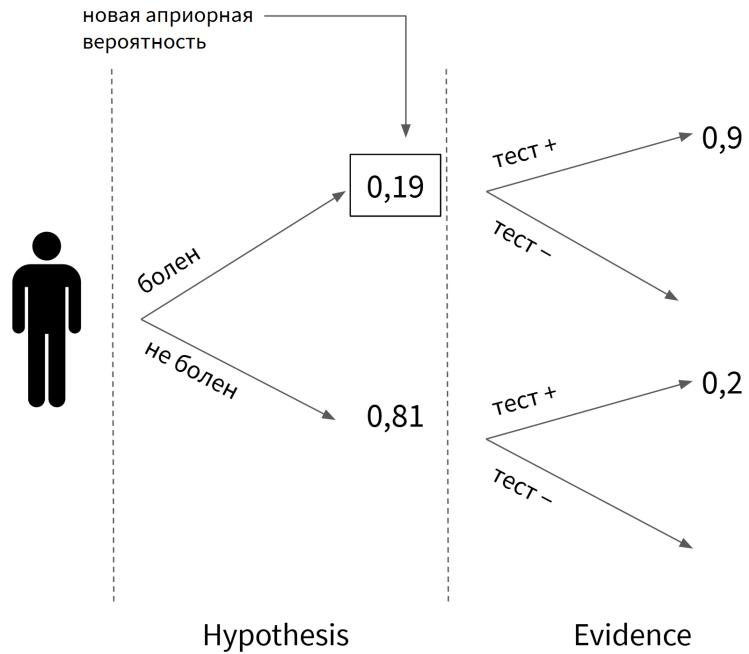


Рис. 31: Новая априорная вероятность

```

prior = np.array([0.05, 0.95])
likelihood = np.array([0.9, 0.2])

for i in range(3):
    posterior = prior * likelihood / np.dot(prior, likelihood)
    print(posterior)
    prior = posterior

print()
print(posterior[0])

Output:
[0.19148936 0.80851064]
[0.51592357 0.48407643]
[0.82746879 0.17253121]

0.8274687854710556
    
```

Листинг 9: Обновление вероятности с помощью цикла

С другой стороны, посмотрим как на схеме выглядит получение повторного положительного теста (рисунок 32).

Обратим внимание, что для вычисления вероятности быть больным после двух положительных результатов достаточно возвести вероятности соответствующих «ветвей» в степень, равную количеству тестов.

$$P(H | E) = \frac{P(H) \cdot P(E | H)^n}{P(E)}$$

$$P(H | E) = \frac{P(H) \cdot P(E | H)^n}{P(H) \cdot P(E | H)^n + P(\neg H) \cdot P(E | \neg H)^n}$$

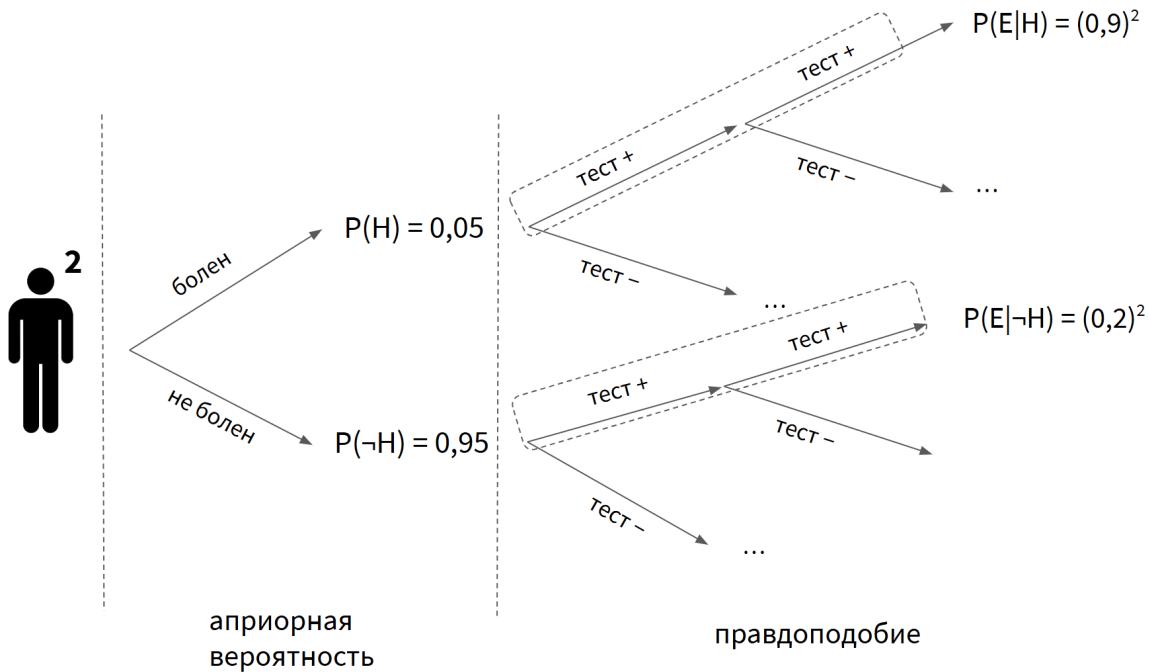


Рис. 32: Вероятность заболеть при повторном положительном тесте

Таким образом, проведение n тестов можно выразить так:

$$\text{posterior}_n = \frac{\text{prior} \times \text{likelihood}^n}{\text{data}}$$

$$\text{posterior}_n \propto \text{prior} \times \text{likelihood}^n$$

Обратимся к Питону (листинг 10).

```
prior = np.array([0.05, 0.95])
likelihood = np.array([0.9, 0.2])

new_posterior = prior[0] * likelihood[0] ** 2 / (prior[0] * likelihood[0] ** 2 + prior[1] * likelihood[1] ** 2)
new_posterior

Output: 0.5159235668789809
```

Листинг 10: Апостериорное распределение как степень правдоподобия

Кроме этого, можно воспользоваться скалярным произведением (листинг 11).

```
prior[0] * likelihood[0] ** 2 / np.dot(prior, likelihood ** 2)

Output: 0.5159235668789809
```

Листинг 11: Апостериорное распределение через скалярное произведение

Посмотрим на вероятность заболеть после пяти положительных тестов (листинг 12).

```

prior[0] * likelihood[0] ** 5 / np.dot(prior, likelihood ** 5)

Output: 0.9898084047136129

```

Листинг 12: Вероятность после пяти положительных тестов

Биномиальное распределение. Несложно заметить, что правдоподобие при однократном teste следует испытаниям Бернулли, при двух и более результатах — биномиальным распределениям для случаев, когда человек болен и когда здоров (рисунок 33).

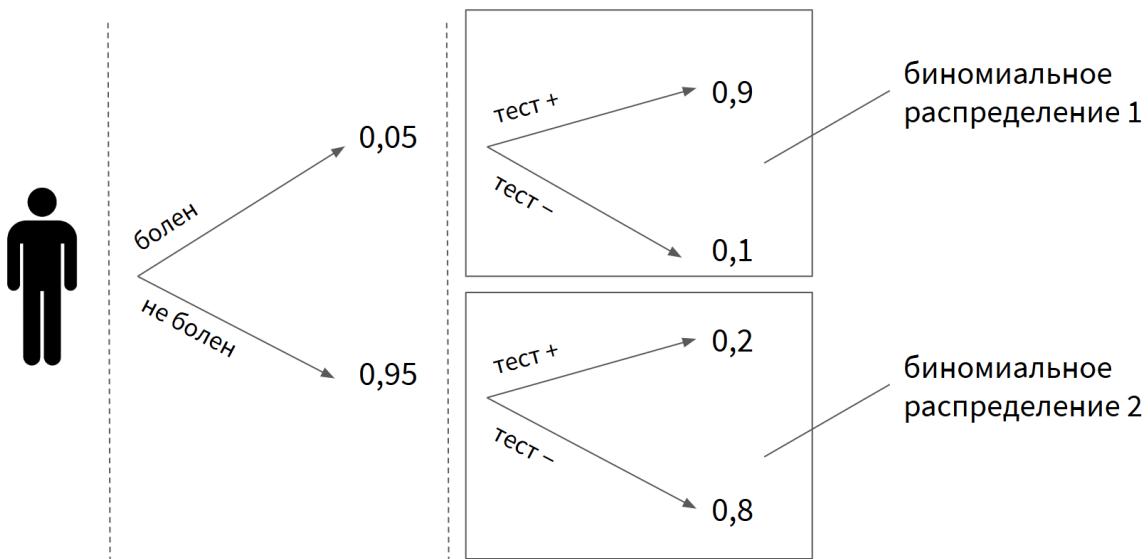


Рис. 33: Биномиальное распределение

Убедимся в этом с помощью Питона. Смоделируем два испытания Бернулли, т.е. проведем однократный тест (листинг 13).

```

from scipy.stats import binom
binom.pmf(k = 1, n = 1, p = 0.9), binom.pmf(k = 1, n = 1, p = 0.2)

Output: (0.9, 0.2)

```

Листинг 13: Однократный тест как испытания Бернулли

Теперь смоделируем два биномиальных распределения, т.е. найдем результаты после двух тестов (листинг 14).

```

prior = np.array([0.05, 0.95])
likelihood = binom.pmf(k = 2, n = 2, p = 0.9), binom.pmf(k = 2, n = 2,
    p = 0.2)
prior[0] * likelihood[0] / np.dot(prior, likelihood)

Output: 0.5159235668789809

```

Листинг 14: Биномиальное распределение при двух тестах

Примечание. Если проводить аналогию с двукратным подбрасыванием монеты, то в случае каждого из двух распределений мы рассчитываем вероятность выпадения двух решек (HH).

Влияние правдоподобия и априорной вероятности. Предположим, что мы ошиблись с оценкой распространенности заболевания и считаем, что оно чрезвычайно редкое (0,000001 населения). Посмотрим (при той же чувствительности), что будет с условной вероятностью быть больным при получении 1, 2, …, 10 положительных результатов (листинг 15).

```

prior = np.array([0.000001, 1 - 0.000001])

for i in range(1, 11):
    likelihood = binom.pmf(k = i, n = i, p = 0.9), binom.pmf(k = i, n =
        i, p = 0.2)
    print(prior[0] * likelihood[0] / np.dot(prior, likelihood))

Output:
4.4999842500551244e-06
2.0249610195003743e-05
9.111678809947254e-05
0.00040989482739716776
0.0018418842973257723
0.008235389119311678
0.036020987079196466
0.14394659706142984
0.4307448121090599
0.7729886606345894

```

Листинг 15: Влияние правдоподобия и априорной вероятности

Как мы видим, при первых трех тестах вероятность быть больным близка к нулю, однако уже к 10-ому тесту она существенно возрастает.

Отсюда следует важный вывод: получаемые данные (результаты теста) в конечном счете оказываются важнее наших априорных оценок (распространенности заболевания).

Если априорные оценки изначально были «на нашей стороне», то есть подкрепляются получаемыми данными, то мы быстрее придем к высокой условной вероятности события.

Если же нет, то мы все равно сделаем правильный вывод, однако для этого нужно будет собрать больше данных.

Посмотрим на формулу под другим углом.

2.4 Коэффициент Байеса

Вероятность и шансы. Оценить наступление случайного события можно двумя способами: с помощью вероятности (probability, P) и с помощью шансов (odds, O).

Как мы уже знаем, согласно, например, классическому определению вероятность — это отношение благоприятствующих событию A исходов k к количеству всех возможных исходов n испытания.

Например, выпадению двойки или тройки на игральной кости благоприятствуют два исхода, $k = 2$, а всего возможных исходов шесть, $n = 6$. Тогда,

$$P(A) = \frac{k}{n} = \frac{2}{6} = \frac{1}{3}$$

Эту же неопределенность можно выразить с помощью шансов на то, что выпадет двойка или тройка. Шансы определяются как отношение благоприятствующих событий исходов к неблагоприятствующим.

В примере с игральной костью шансы выпадения этих значений составят два к четырем или, если сократить на два, один к двум (рисунок 34).

$$O(A) = 2 : 4 = 1 : 2$$

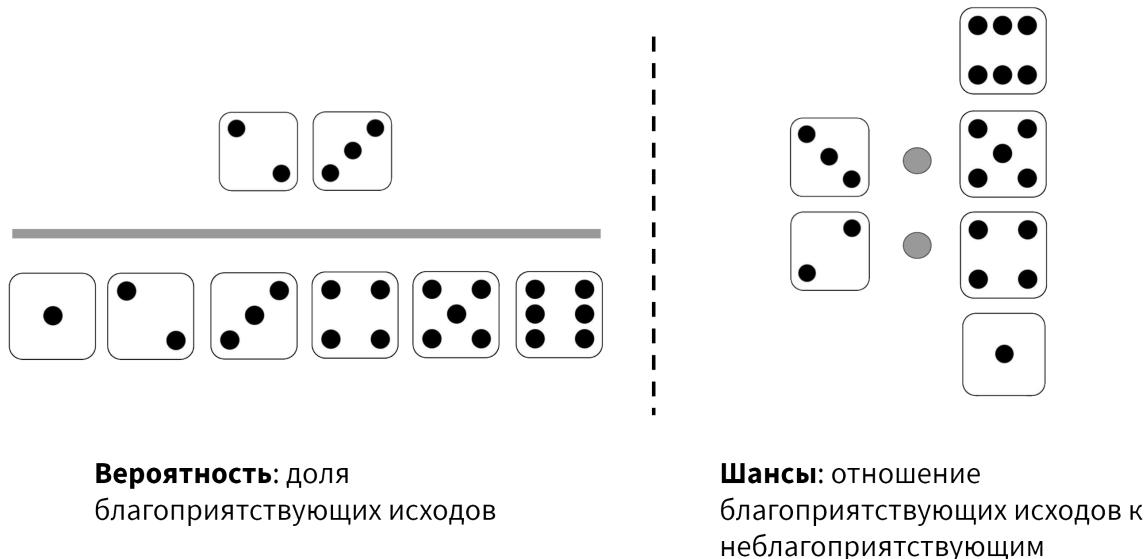


Рис. 34: Вероятность и шансы

Интуитивно можно сказать, что такое отношение показывает, насколько благоприятствующие исходы более (или менее) значимы, чем неблагоприятствующие. В примере выше, неблагоприятствующие исходы в два раза более значимы.

Отношение правдоподобия. Эту же идею выражения неопределенности через шансы можно применить и к формуле Байеса. В примере с медицинскими тестами (рисунок 35):

- условно «благоприятствующими» исходами для вероятности быть больным при положительном teste будут те случаи, когда пациент получил положительный тест при условии, что он болен;
- «неблагоприятствующими» исходами будут наоборот ситуации, когда тест положителен, а человек не болен.

Получается, что речь идет об **отношении правдоподобия** (likelihood ratio, Λ).

$$\Lambda_{+ \mid \text{болен}} = \frac{P(B \mid A_1)}{P(B \mid A_2)} = \frac{0,9}{0,2} = 4,5$$

Отношение правдоподобия можно также выразить как шансы быть больным при положительном teste

$$O_{+ \mid \text{болен}} = B \mid A_1 : B \mid A_2 = 9/10 : 2/10 = 9 : 2 = 4,5 : 1$$

Другими словами, при положительном teste шансы быть больным в 4,5 раза больше, чем быть здоровым. Число 4,5, в данном случае, называется **коэффициентом Байеса** (Bayes factor, BF).

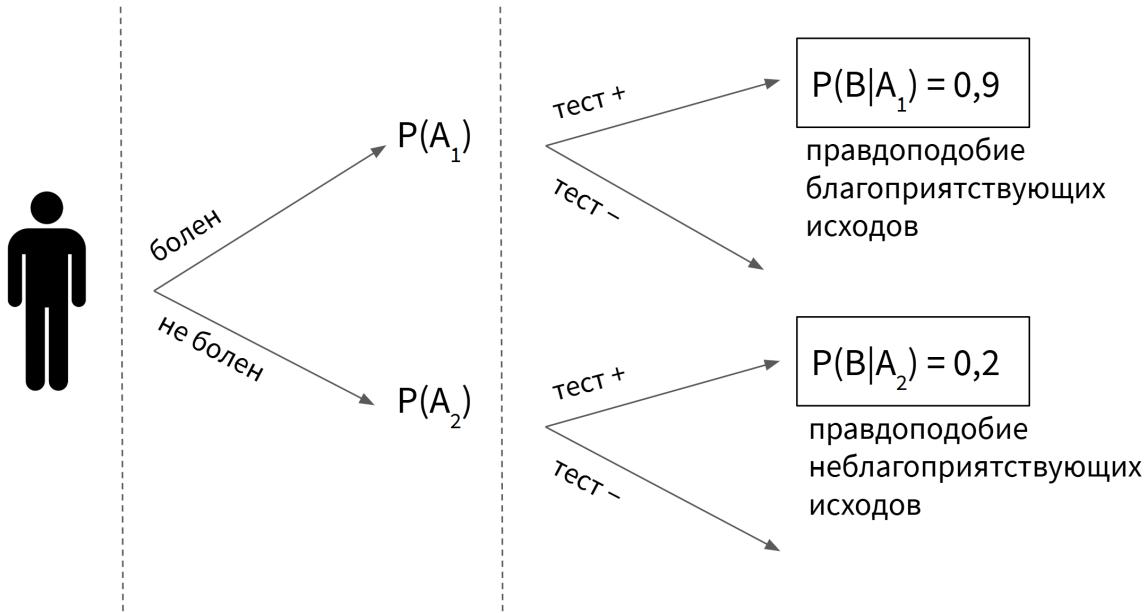


Рис. 35: Отношение правдоподобия

Шансы быть больным при положительном тесте. Отношение правдоподобия или коэффициент Байеса также позволяют обновить априорную вероятность. Единственный нюанс — ее также необходимо выразить в виде опять же априорных шансов (prior odds).

$$O_{\text{болен} | +} = O_{\text{болен}} \cdot O_{+ | \text{болен}} = (5 : 95) \times (4,5 : 1) = 22,5 : 95$$

Преобразуем шансы обратно в вероятность.

$$P(\text{болен} | +) = \frac{22,5}{22,5 + 95} \approx 0,19$$

На уже знакомой схеме (рисунок 36) априорные шансы и отношение правдоподобия можно выразить так:

$$\begin{aligned} \text{априорные шансы} & \quad \text{отношение} \\ & \quad \text{правдоподобия} \\ & \quad (\text{коэффициент Байеса}) \\ P(\text{болен} | +) & \propto [0,05] \cdot [0,90] = 0,045 \\ P(\text{не болен} | +) & \propto [0,95] \cdot [0,20] = 0,19 \end{aligned}$$

Рис. 36: Априорные шансы и отношение правдоподобия

Найдем также шансы быть здоровым при положительном тесте.

$$O_{\text{не болен} | +} = O_{\text{не болен}} \cdot O_{+ | \text{не болен}} = (95 : 5) \times (1 : 4,5) = 95 : 22,5$$

Тогда вероятность

$$P(\text{не болен} | +) = \frac{95}{95 + 22,5} \approx 0,81$$

Почему это работает. Продемонстрируем графически (рисунок 37), почему умножение априорных шансов на коэффициент Байеса дает те же апостериорные шансы (апостериорную вероятность), что и формула Байеса.

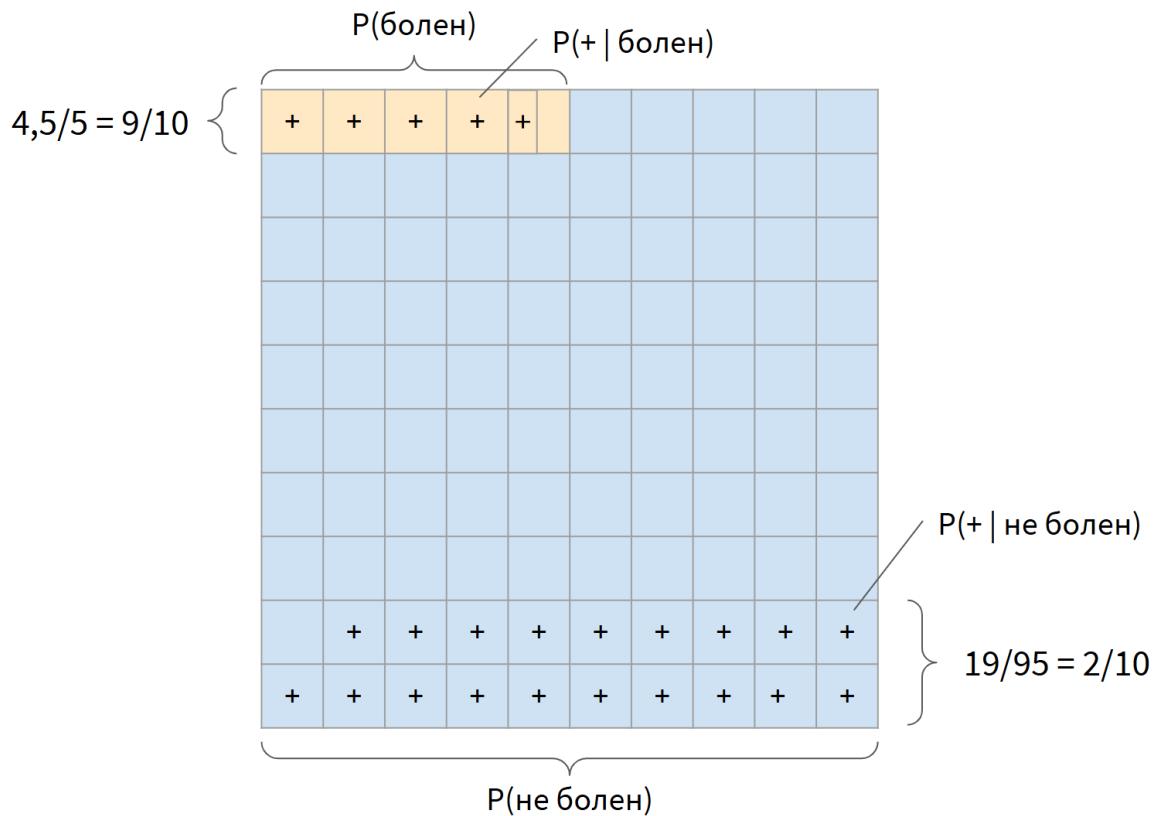


Рис. 37: Апостериорные шансы и апостериорная вероятность

3 Примеры

Рассмотрим несколько примеров для закрепления пройденного материала, а также введем новые термины и понятия.

3.1 Колоды карт

Перед нами две колоды карт (deck of cards):

- одна полная колода из 52-х карт D_{full} ; и
- одна сокращенная, состоящая только из «красных мастей» (черви и бубны) D_{red} и, как следствие, 26-ти карт.

Наугад достанем карту из одной из колод (из какой именно мы не знаем). Пусть это будет тройка червей, $3d$, от англ. three of diamonds (рисунок 38).

Какова вероятность, что тройка червей лежала в полной колоде? Другими словами, нам нужно найти $P(D_{\text{full}} | 3d)$. Начнем с априорной вероятности, то есть выбора колоды. Колода выбирается случайно, а значит

$$P(D_{\text{full}}) = P(D_{\text{red}}) = \frac{1}{2}$$

Далее необходимо найти условные вероятности (правдоподобие) выбора тройки червей из каждой колоды. Вполне очевидно, что вероятность выбрать такую карту из полной

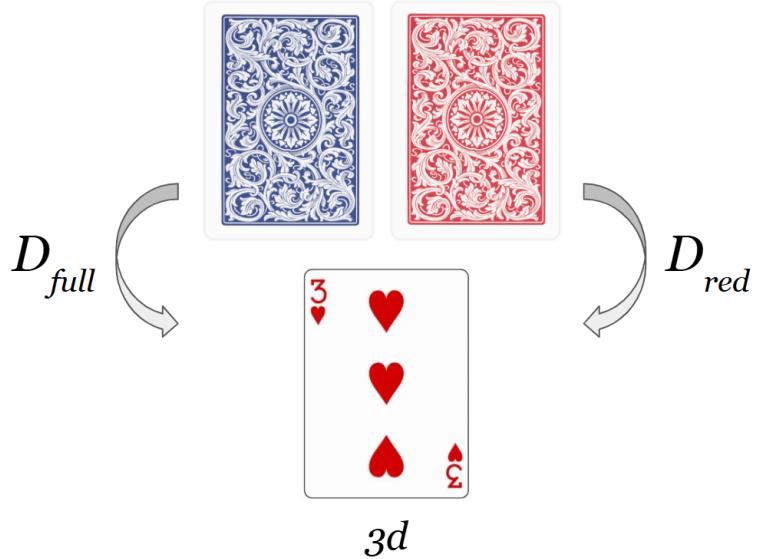


Рис. 38: В какой колоде лежала тройка червей?

колоды ниже, поскольку в ней больше карт.

$$P(3d | D_{\text{full}}) = \frac{1}{52}$$

$$P(3d | D_{\text{red}}) = \frac{1}{26}$$

Теперь по формуле полной вероятности найдем вероятность выпадения тройки червей $P(3d)$. Напомню, это сумма совместных вероятностей

- выбора полной колоды и условной вероятности достать из нее тройку червей, $P(D_{\text{full}}) \cdot P(3d | D_{\text{full}})$; и
- выбора красной колоды и условной вероятности достать из нее ту же карту, $P(D_{\text{red}}) \cdot P(3d | D_{\text{red}})$.

Таким образом,

$$P(3d) = P(D_{\text{full}}) \cdot P(3d | D_{\text{full}}) + P(D_{\text{red}}) \cdot P(3d | D_{\text{red}}) = \frac{1}{2} \cdot \frac{1}{52} + \frac{1}{2} \cdot \frac{1}{26} = \frac{3}{104}$$

Подставим значения в формулу Байеса.

$$P(D_{\text{full}} | 3d) = \frac{P(D_{\text{full}}) \cdot P(3d | D_{\text{full}})}{P(3d)} = \frac{\frac{1}{2} \cdot \frac{1}{52}}{\frac{3}{104}} = \frac{1}{3}$$

Такой результат вполне логичен. В полной колоде в два раза больше карт, чем в красной, а значит и вероятность того, что это именно полная колода, должна быть в два раза меньше.

Если перевести вероятности в шансы, то получим

$$O(D_{\text{full}} | 3d) : O(D_{\text{red}} | 3d) = \frac{1}{3} : \frac{2}{3} = 1 : 2$$

И коэффициент Байеса равен

$$BF_{D_{\text{full}}|3d} = \frac{P(3d | D_{\text{full}})}{P(3d | D_{\text{red}})} = \frac{1/52}{1/26} = \frac{1}{2}$$

Еще раз поясним словами. Коэффициент Байеса или отношение правдоподобия того, что тройка червей взята из полной колоды равен 0,5, а значит шансы такого действия равны $\frac{1}{2} : 1$ или (умножив на 2) $1 : 2$, что мы и получили выше.

При умножении коэффициента Байеса, выраженного в шансах, на априорные шансы, мы получим апостериорные шансы. Так как априорная вероятность для каждой колоды равна $\frac{1}{2}$, то шансы будут $1 : 1$, а значит не повлияют на апостериорные шансы.

$$O(D_{\text{full}} \mid 3d) = (1 : 1) \times (1 : 2) = 1 : 2$$

При переводе в вероятность получим $P(D_{\text{full}} \mid 3d) = \frac{1}{1+2} = \frac{1}{3}$.

3.2 Игровые кости

Вместо колоды карт рассмотрим три игровые кости (dice), имеющие четыре (D_1), шесть (D_2) и восемь (D_3) граней (рисунок 39).

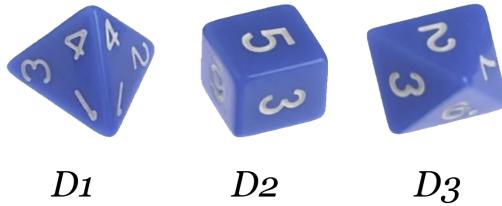


Рис. 39: Игровые кости

Предположим, что кто-то выбирает одну из костей и бросает три раза, каждый раз сообщая нам результат броска, но не показывая саму кость. Наша задача — после трех бросков догадаться, какая именно кость была брошена.

Расчет вручную. Изначально мы ничего не знаем о выбранной кости, поэтому наши априорные вероятности будут одинаковыми.

$$P(D_1) = P(D_2) = P(D_3) = \frac{1}{3}$$

Попросим нашего помощника бросить кость.

Первый бросок. Предположим, что в первый раз выпала четверка. Обозначим этот факт как $r4$ (от англ. to roll 4 on a dice). Найдем правдоподобие такого результата для каждой из костей:

- на первой кости четыре грани, значит $P(r4 \mid D_1) = \frac{1}{4}$;
- на второй кости шесть, $P(r4 \mid D_2) = \frac{1}{6}$;
- на третьей — восемь, $P(r4 \mid D_3) = \frac{1}{8}$.

Также найдем полную вероятность выпадения четверки или данные, $P(r4)$. Она будет состоять из суммы совместных вероятностей выпадения четверки на каждой из костей.

$$P(r4) = P(D_1) \cdot P(r4 \mid D_1) + P(D_2) \cdot P(r4 \mid D_2) + P(D_3) \cdot P(r4 \mid D_3) =$$

$$\frac{1}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{8} = \frac{13}{72}$$

Остается найти апостериорные вероятности.

$$P(D1 | r4) = \frac{P(D1) \cdot P(r4 | D1)}{P(r4)} = \frac{\frac{1}{3} \cdot \frac{1}{4}}{\frac{13}{72}} = \frac{6}{13} \approx 0,46$$

$$P(D2 | r4) = \frac{P(D2) \cdot P(r4 | D2)}{P(r4)} = \frac{\frac{1}{3} \cdot \frac{1}{6}}{\frac{13}{72}} = \frac{4}{13} \approx 0,31$$

$$P(D3 | r4) = \frac{P(D3) \cdot P(r4 | D3)}{P(r4)} = \frac{\frac{1}{3} \cdot \frac{1}{8}}{\frac{13}{72}} = \frac{3}{13} \approx 0,23$$

Опять же, такой результат вполне ожидаем. На первой кости только четыре грани и вероятность (правдоподобие) выпадения значений от одного до четырех здесь наиболее высокая.

Второй бросок. Пусть в этот раз выпадет двойка. Апостериорная вероятность станет новой априорной.

$$P(D1 | r4) = \frac{6}{13}, \quad P(D2 | r4) = \frac{4}{13}, \quad P(D3 | r4) = \frac{3}{13}$$

Правдоподобие будет тем же.

$$P(r2 | D1) = \frac{1}{4}, \quad P(r2 | D2) = \frac{1}{6}, \quad P(r2 | D3) = \frac{1}{8}$$

Полная вероятность $P(r2)$ теперь должна учитывать новую априорную вероятность.

$$P(r2) = P(D1 | r4) \cdot P(r2 | D1) + P(D2 | r4) \cdot P(r2 | D2) + P(D3 | r4) \cdot P(r2 | D3) = \\ \frac{6}{13} \cdot \frac{1}{4} + \frac{4}{13} \cdot \frac{1}{6} + \frac{3}{13} \cdot \frac{1}{8} = \frac{61}{312}$$

Найдем апостериорные вероятности при втором броске.

$$P(D1 | r2) = \frac{P(D1 | r4) \cdot P(r2 | D1)}{P(r2)} = \frac{\frac{6}{13} \cdot \frac{1}{4}}{\frac{61}{312}} = \frac{36}{61} \approx 0,59$$

$$P(D2 | r2) = \frac{P(D2 | r4) \cdot P(r2 | D2)}{P(r2)} = \frac{\frac{4}{13} \cdot \frac{1}{6}}{\frac{61}{312}} = \frac{16}{61} \approx 0,26$$

$$P(D3 | r2) = \frac{P(D3 | r4) \cdot P(r2 | D3)}{P(r2)} = \frac{\frac{3}{13} \cdot \frac{1}{8}}{\frac{61}{312}} = \frac{9}{61} \approx 0,15$$

Так как правдоподобие двойки на первой кости снова самое высокое, мы только укрепились во мнении, что речь идет о первой кости.

Третий бросок. При третьем броске выпадает пятерка. Найдем апостериорные вероятности.

$$P(D1 | r5) = \frac{P(D1 | (r4, r2)) \cdot P(r5 | D1)}{P(r5)} = \frac{\frac{36}{61} \cdot 0}{\frac{91}{1464}} = 0$$

$$P(D2 | r5) = \frac{P(D2 | (r4, r2)) \cdot P(r5 | D2)}{P(r5)} = \frac{\frac{16}{61} \cdot \frac{1}{6}}{\frac{91}{1464}} = \frac{64}{91} \approx 0,70$$

$$P(D3 | r5) = \frac{P(D3 | (r4, r2)) \cdot P(r5 | D3)}{P(r5)} = \frac{\frac{9}{61} \cdot \frac{1}{8}}{\frac{91}{1464}} = \frac{27}{91} \approx 0,30$$

Обратим внимание, что правдоподобие выпадения пяты на первой кости равно нулю (такой грани там нет), а значит все это время речь шла о второй или третьей костях.

Более того, так как новая априорная вероятность также равна нулю, вне зависимости от того, какая грань выпадет следующей (пусть даже от 1 до 4), вероятность первой кости всегда будет нулевой.

Из оставшихся двух вариантов на такую высокую вероятность второй кости повлияло два фактора:

- во-первых, выпадение пятерки на ней более правдоподобно, $\frac{1}{6}$ против $\frac{1}{8}$;
- во-вторых, за первые два броска накопилась более высокая априорная вероятность, $\frac{16}{61}$ против $\frac{9}{61}$.

Методы оценки. Оценка в данном случае означает выбор одной из гипотез или, в нашем примере, одной из трех костей. Еще на первом занятии мы сказали, что знаменатель здесь роли не играет. Остается два компонента: априорная вероятность и правдоподобие.

$$P(\theta | X) = \frac{P(\theta) \cdot P(X | \theta)}{P(X)}$$

$$P(\theta | X) \propto P(\theta) \cdot P(X | \theta)$$

Апостериорный максимум. Если выбирать гипотезу исходя из обоих компонентов, то мы по сути основываемся на оценке с помощью **апостериорного максимума** (maximum a posteriori, MAP). После трех бросков максимальная апостериорная вероятность у второй кости, ее мы и выбираем (рисунок 40).

$$\theta_{MAP} = D_2$$

В общем виде: $\theta_{MAP} = \operatorname{argmax}_{\theta} P(\theta) \cdot P(X | \theta)$

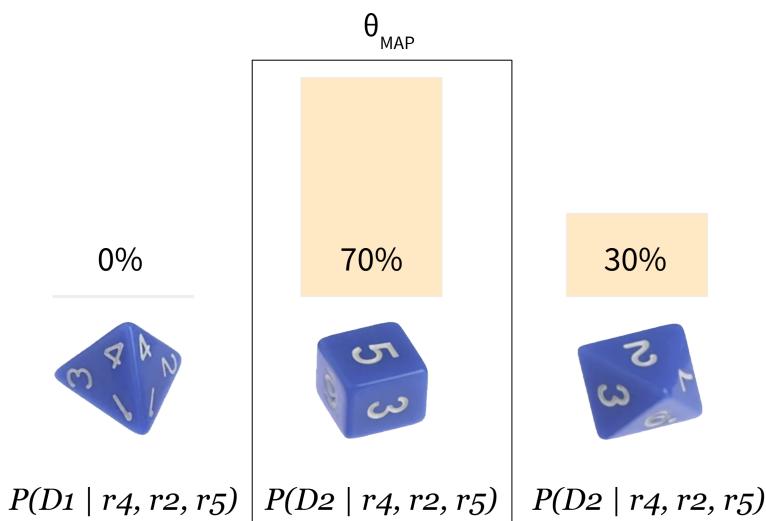


Рис. 40: Апостериорный максимум

Примечание. **Аргумент максимизации** (argument of the maxima)

$$\operatorname{argmax}_x f(x)$$

— это такое значение x , при котором $f(x)$ принимает наибольшее значение.

Метод максимального правдоподобия. Если же не учитывать априорную вероятность и оценивать θ только через правдоподобие, то наиболее вероятной в первых двух бросках является первая кость $\frac{1}{4}$, а при третьем броске — вторая $\frac{1}{6}$.

Посмотрим на оценку правдоподобия после первого броска (рисунок 41).

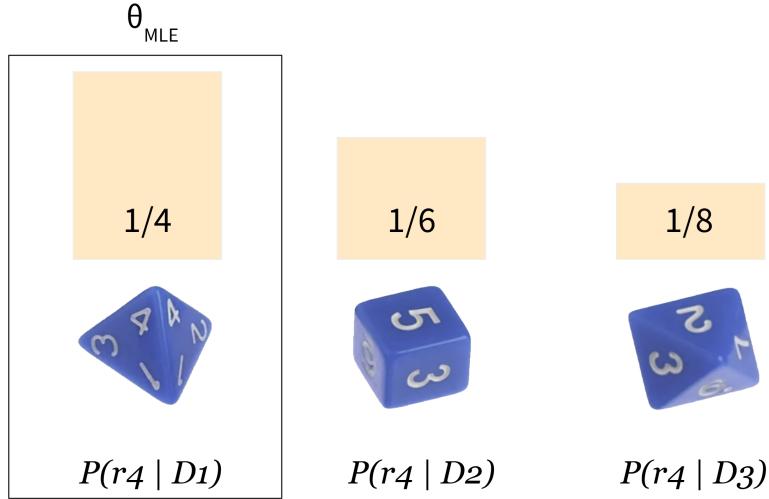


Рис. 41: Метод максимального правдоподобия

В этом случае говорят, что мы оцениваем параметр с помощью **метода максимального правдоподобия** (maximum likelihood estimation, MLE).

$$\theta_{MLE} = \operatorname{argmax}_{\theta} P(X | \theta)$$

Интересно, что если априорная вероятность для каждой из гипотез или для каждого θ одинакова (можно считать, что она представляет собой константу), то обе оценки дают одинаковый результат.

Формально,

$$\theta_{MAP} = \operatorname{argmax}_{\theta} const \cdot P(X | \theta) = \operatorname{argmax}_{\theta} P(X | \theta) = \theta_{MLE}$$

Так было, в частности, после первого броска кости. Так как все априорные вероятности были одинаковыми и равнялись $\frac{1}{3}$, то единственным, что влияло на результат было правдоподобие. Приведем расчеты еще раз.

$$\begin{aligned} P(D1 | r4) &= \frac{P(D1) \cdot P(r4 | D1)}{P(r4)} = \frac{\frac{1}{3} \cdot \frac{1}{4}}{\frac{13}{72}} = \frac{6}{13} \approx 0,46 \\ P(D2 | r4) &= \frac{P(D2) \cdot P(r4 | D2)}{P(r4)} = \frac{\frac{1}{3} \cdot \frac{1}{6}}{\frac{13}{72}} = \frac{4}{13} \approx 0,31 \\ P(D3 | r4) &= \frac{P(D3) \cdot P(r4 | D3)}{P(r4)} = \frac{\frac{1}{3} \cdot \frac{1}{8}}{\frac{13}{72}} = \frac{3}{13} \approx 0,23 \end{aligned}$$

Максимальным правдоподобие было у первой кости $\frac{1}{4} = 0,25$ (оценка методом максимального правдоподобия) и именно эта кость показала наибольшую апостериорную вероятность $\frac{6}{13} \approx 0,46$ (оценка с помощью апостериорного максимума).

В случае если априорная вероятность и правдоподобие существенно расходятся, МАР и МЛЕ-оценки могут не совпадать.

Таким образом, оценка методом максимального правдоподобия является частным случаем оценки с помощью апостериорного максимума при одинаковой априорной вероятности каждой из гипотез.

Расчет с помощью Питона. Расчет апостериорной вероятности вручную довольно кропотлив. Доверим эту работу компьютеру.

Вначале повторим те результаты, которые мы получили в примере выше.

Три броска. Начнем с одного броска и выпадения четверки (листинг 16).

```
likelihood = np.array([1/4, 1/6, 1/8])
prior = np.array([1/3, 1/3, 1/3])

(likelihood * prior) / np.dot(likelihood, prior)

Output: array([0.46153846, 0.30769231, 0.23076923])
```

Листинг 16: Один бросок и выпадение четверки

Теперь выполним три броска с выпадением четверки, двойки и пятерки (листинг 17).

```
prior = np.array([1/3, 1/3, 1/3])
rolls = np.array([4, 2, 5])

for r in rolls:
    likelihood = np.array([1/4, 1/6, 1/8])
    if r > 4:
        likelihood[0] = 0
    if r > 6:
        likelihood[1] = 0

    posterior = (likelihood * prior) / np.dot(likelihood, prior)
    print(posterior)
    prior = posterior

Output:
[0.46153846 0.30769231 0.23076923]
[0.59016393 0.26229508 0.14754098]
[0.          0.7032967   0.2967033]
```

Листинг 17: Три броска и выпадение четверки, двойки и пятерки

МАР и МЛЕ. Найдем оценки МАР и МЛЕ для случаев одинаковой и разной априорной вероятности при выпадении четверки. Случай одинаковой априорной оценки полностью повторяет первый бросок в примере выше (листинг 18).

```

likelihood = np.array([1/4, 1/6, 1/8])
prior = np.array([1/3, 1/3, 1/3])
posterior = (likelihood * prior) / np.dot(likelihood, prior)
MAP = np.argmax(posterior) + 1
MLE = np.argmax(likelihood) + 1

posterior, MAP, MLE

Output: (array([0.46153846, 0.30769231, 0.23076923]), 1, 1)

```

Листинг 18: Одинаковая априорная вероятность

Если предположить, что априорная вероятность первой кости меньше остальных (например, $\frac{1}{5}$ против $\frac{2}{5}$ у второй и третьей кости), то MAP и MLE не совпадут (листинг 19).

```

likelihood = np.array([1/4, 1/6, 1/8])
prior = np.array([1/5, 2/5, 2/5])

posterior = (likelihood * prior) / np.dot(likelihood, prior)
MAP = np.argmax(posterior) + 1
MLE = np.argmax(likelihood) + 1

posterior, MAP, MLE

Output: (array([0.3, 0.4, 0.3]), 2, 1)

```

Листинг 19: Разная априорная вероятность

Отгадывание кости. Поместим код выше в функцию, которая будет оценивать уже четыре вида игральных костей с четырьмя, шестью, восьмью и двенадцатью гранями и после произвольного числа бросков отгадывать, о какой кости идет речь.

Такого рода программы, которые последовательно исследуют «мир» (world) и делают выводы называют **интеллектуальными агентами** (intelligent agent) или просто агентами.

В нашем примере мир — это результаты бросания одной из четырех костей, а вывод — MAP-оценка того, о какой кости идет речь (листинг 20).

```

def dice_agent(prior, rolls):

    for r in rolls:
        likelihood = np.array([1/4, 1/6, 1/8, 1/12])
        if r > 4:
            likelihood[0] = 0
        if r > 6:
            likelihood[1] = 0
        if r > 8:
            likelihood[2] = 0

        posterior = (likelihood * prior) / np.dot(likelihood, prior)
        prior = posterior

    return posterior, np.argmax(posterior) + 1

```

Листинг 20: Функция интеллектуального агента

Бросим первую кость с четырьмя гранями пять раз (листинг 21).

```

prior = np.array([1/4, 1/4, 1/4, 1/4])

np.random.seed(42)
rolls = np.random.randint(1, 5, size = 5)

dice_agent(prior, rolls)

Output: (array([0.8568595 , 0.11283747, 0.02677686, 0.00352617]), 1)

```

Листинг 21: Пять бросков кости с четырьмя гранями

Усложним задачу. Будем бросать вторую кость только три раза и изменим априорную вероятность не в пользу этого варианта (листинг 22).

```

prior = np.array([1/3, 1/6, 1/6, 1/3])

np.random.seed(42)
rolls = np.random.randint(1, 7, size = 3)

dice_agent(prior, rolls)

Output: (array([0.          , 0.59813084, 0.25233645, 0.14953271]), 2)

```

Листинг 22: Три броска кости с шестью гранями

Как мы видим, агент в обоих случаях успешно справился с поставленной задачей.

4 Оценка распределения

Выше мы как в теории, так и на практике познакомились с формулой Байеса и научились находить апостериорную вероятность гипотез с учетом полученных данных.

Теперь давайте исследуем, что будет, если:

- формулировать больше гипотез или значений θ (например, увеличивать количество игральных костей); и одновременно
- собирать больше данных (то есть увеличивать количество бросков).

Для удобства изменим фабулу задачи и предположим, что перед нами не кости, а мешки, из которых мы достаем шары двух цветов, белого и черного.

4.1 Пять мешков

Начнем с *пяти* мешков. Зададим априорную вероятность. Она будет одинаковой для всех мешков и соответственно равна $\frac{1}{5}$ (листинг 23).

```
initial_prior = 1/5
```

Листинг 23: Априорная вероятность

Пусть правдоподобие появления *белого* шара равномерно увеличивается от нуля для первого мешка до единицы для пятого (листинг 24).

```
white_likelihood = np.linspace(0, 1, num = 5)
white_likelihood
```

```
Output: array([0.    , 0.25, 0.5   , 0.75, 1.    ])
```

Листинг 24: Правдоподобие белого шара

Заметим при этом, что правдоподобие появления *черного* шара (листинг 25) будет в каждом случае равно (1 – правдоподобие появления белого шара).

```
black_likelihood = 1 - white_likelihood
black_likelihood
```

```
Output: array([1.    , 0.75, 0.5   , 0.25, 0.    ])
```

Листинг 25: Правдоподобие черного шара

Поставим задачу найти апостериорную вероятность появления одного белого шара из каждого из пяти мешков (рисунок 42).

Один белый шар. Апостериорную вероятность для относительно большого количества мешков удобно найти с помощью таблицы. Создадим таблицу и найдем числитель формулы Байеса (листинг 26 и рисунок 43).

```
w = pd.DataFrame(index = ['B1', 'B2', 'B3', 'B4', 'B5'])

w['prior'] = initial_prior
w['likelihood'] = white_likelihood
w['numerator'] = w['prior'] * w['likelihood']

w
```

Листинг 26: Вероятность появления одного белого шара. Числитель

?	?	?	?	?	апостериорная вероятность каждого из мешков при условии одного белого шара
<input type="radio"/>	результат первого испытания				
1	0,75	0,5	0,25	0	правдоподобие черного шара
0	0,25	0,5	0,75	1	правдоподобие белого шара
B1 $\frac{1}{5}$	B2 $\frac{1}{5}$	B3 $\frac{1}{5}$	B4 $\frac{1}{5}$	B5 $\frac{1}{5}$	априорная вероятность каждого из мешков

Рис. 42: Апостериорная вероятность появления белого шара

	prior	likelihood	numerator
B1	0.2	0.00	0.00
B2	0.2	0.25	0.05
B3	0.2	0.50	0.10
B4	0.2	0.75	0.15
B5	0.2	1.00	0.20

Рис. 43: Вероятность появления одного белого шара. Числитель

Найдем знаменатель, то есть вероятность данных. Для этого просуммируем все возможные совместные вероятности нашего априорного знания о мешках и правдоподобия соответствующего мешка (листинг 27).

```
prob_data = w['numerator'].sum()
prob_data

Output: 0.5
```

Листинг 27: Вероятность появления одного белого шара. Знаменатель

Найдем апостериорную вероятность (листинг 28 и рисунок 44).

```
w['posterior'] = w['numerator'] / prob_data
w
```

Листинг 28: Апостериорная вероятность появления одного белого шара

Обратим внимание, что сумма апостериорных вероятностей (как и априорных) всегда равна единице (листинг 29).

	prior	likelihood	numerator	posterior
B1	0.2	0.00	0.00	0.0
B2	0.2	0.25	0.05	0.1
B3	0.2	0.50	0.10	0.2
B4	0.2	0.75	0.15	0.3
B5	0.2	1.00	0.20	0.4

Рис. 44: Апостериорная вероятность появления одного белого шара

```
w['posterior'].sum()
Output: 1.0
```

Листинг 29: Сумма апостериорных вероятностей

При этом сумма правдоподобий единице равна быть не должна (листинг 30). О том, почему это так, мы поговорим чуть позже.

```
w['likelihood'].sum()
Output: 2.5
```

Листинг 30: Сумма правдоподобий

Посмотрим на апостериорную вероятность на графике (листинг 31 и рисунок 45).

```
w['posterior'].plot.bar();
```

Листинг 31: Распределение апостериорной вероятности появления одного белого шара

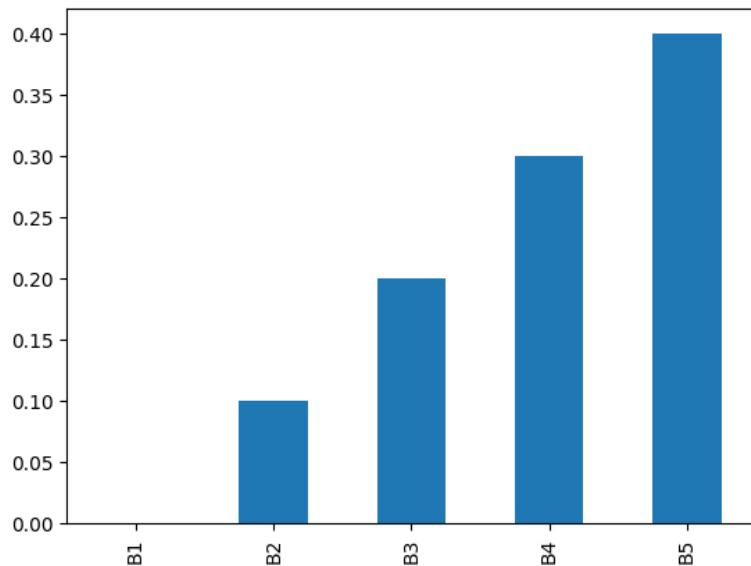


Рис. 45: Распределение апостериорной вероятности появления одного белого шара

Разумеется, при наибольшем правдоподобии мешка B5 и одинаковой априорной вероят-

ности именно этот мешок показывает максимальную апостериорную оценку.

Два белых шара. По аналогичной схеме найдем апостериорную вероятность появления двух белых шаров (листинг 32 и рисунок 46).

```
ww = pd.DataFrame(index = ['B1', 'B2', 'B3', 'B4', 'B5'])
ww['prior'] = initial_prior
ww['likelihood'] = white_likelihood ** 2
ww['numerator'] = ww['prior'] * ww['likelihood']

prob_data = ww['numerator'].sum()

ww['posterior'] = ww['numerator'] / prob_data
ww
```

Листинг 32: Апостериорная вероятность появления двух белых шаров

	prior	likelihood	numerator	posterior
B1	0.2	0.0000	0.0000	0.000000
B2	0.2	0.0625	0.0125	0.033333
B3	0.2	0.2500	0.0500	0.133333
B4	0.2	0.5625	0.1125	0.300000
B5	0.2	1.0000	0.2000	0.533333

Рис. 46: Апостериорная вероятность появления двух белых шаров

Посмотрим на результат на графике (листинг 33 и рисунок 47).

```
ww['posterior'].plot.bar();
```

Листинг 33: Распределение апостериорной вероятности двух белых шаров

После появления двух белых шаров мы только укрепляемся во мнении, что речь идет о мешке B5.

Два белых и один черный шар. Достанем черный шар (листинг 34 и рисунок 48).

```
wwb = pd.DataFrame(index = ['B1', 'B2', 'B3', 'B4', 'B5'])
wwb['prior'] = initial_prior
wwb['likelihood'] = white_likelihood ** 2 * black_likelihood
wwb['numerator'] = wwb['prior'] * wwb['likelihood']

prob_data = wwb['numerator'].sum()

wwb['posterior'] = wwb['numerator'] / prob_data
wwb
```

Листинг 34: Апостериорная вероятность двух белых и одного черного шара

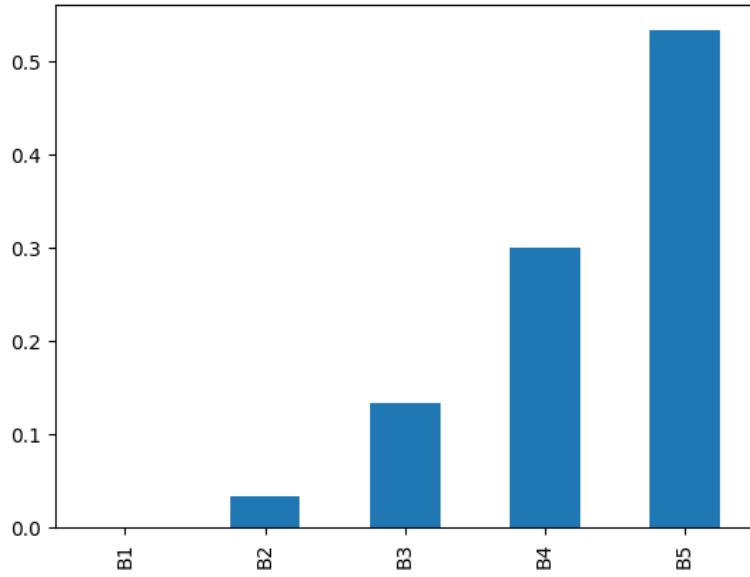


Рис. 47: Распределение апостериорной вероятности двух белых шаров

	prior	likelihood	numerator	posterior
B1	0.2	0.000000	0.000000	0.00
B2	0.2	0.046875	0.009375	0.15
B3	0.2	0.125000	0.025000	0.40
B4	0.2	0.140625	0.028125	0.45
B5	0.2	0.000000	0.000000	0.00

Рис. 48: Апостериорная вероятность двух белых и одного черного шара

Посмотрим на результат на графике (листинг 35 и рисунок 49).

```
wwb['posterior'].plot.bar();
```

Листинг 35: Распределение апостериорной вероятности двух белых и одного черного шара

Такой результат также вполне ожидаем. В мешках B1 и B5 согласно их правдоподобию могут появляться только черный (в B1 правдоподобие белого равно нулю) или только белый (в B5 правдоподобие белого равно единице) шары. Последовательность «белый-белый-черный» из этих мешков достать никак не получится.

При этом так как белых шаров больше, чем черных, и наибольшее правдоподобие белого шара из оставшихся вариантов у мешка B4, то именно этот мешок получает наибольшую апостериорную оценку.

Биномиальное распределение. Вспомним, что правдоподобие в данном случае следует биномиальному распределению. Посмотрим на правдоподобие достать один белый шар, т.е. по сути проведем испытание Бернулли (листинг 36).

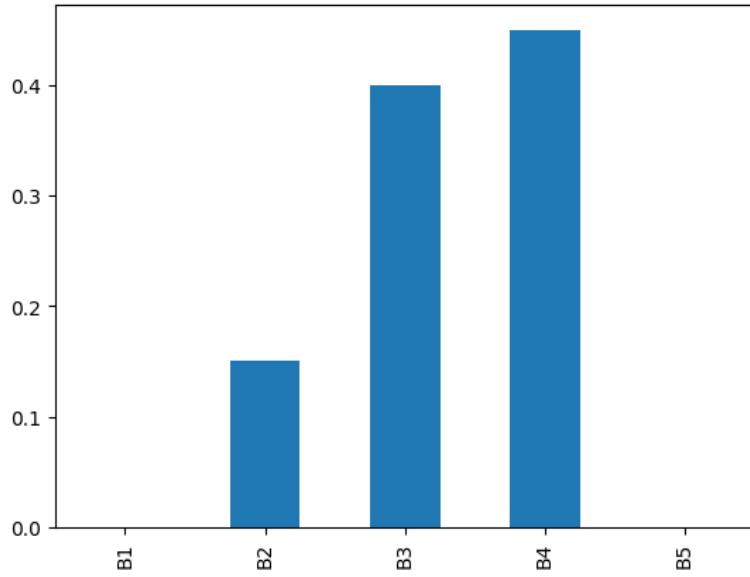


Рис. 49: Распределение апостериорной вероятности двух белых и одного черного шара

```
from scipy.stats import binom
binom.pmf(k = 1, n = 1, p = white_likelihood)

Output: array([0. , 0.25, 0.5 , 0.75, 1. ])
```

Листинг 36: Правдоподобие достать один белый шар

Снова найдем апостериорную вероятность двух белых и одного черного шара (листинг 37 и рисунок 50).

```
wwb2 = pd.DataFrame(index = ['B1', 'B2', 'B3', 'B4', 'B5'])
wwb2['prior'] = initial_prior
wwb2['likelihood'] = binom.pmf(k = 2, n = 3, p = white_likelihood)
wwb2['numerator'] = wwb2['prior'] * wwb2['likelihood']

prob_data = wwb2['numerator'].sum()

wwb2['posterior'] = wwb2['numerator'] / prob_data
wwb2
```

Листинг 37: Биномиальное распределение вероятности двух белых и одного черного шара

Мы видим, что апостериорные вероятности совпадают, а вот правдоподобие и, как следствие, весь числитель — нет. Почему так получается?

Вначале посмотрим на биномиальное распределение для, например, мешка B2 (рисунок 51).

Сравнив столбцы likelihood рассчитанной «вручную» апостериорной вероятности и вероятности, рассчитанной с помощью метода `binom.pmf()`, мы легко убедимся, что они отличаются ровно в три раза, то есть на размер биномиального коэффициента $\binom{3}{2} = 3$ (рисунок 52).

Позднее же, так как биномиальный коэффициент находится в каждом слагаемом знаменателя, его можно сократить с коэффициентом, находящимся в числителе и получить точно

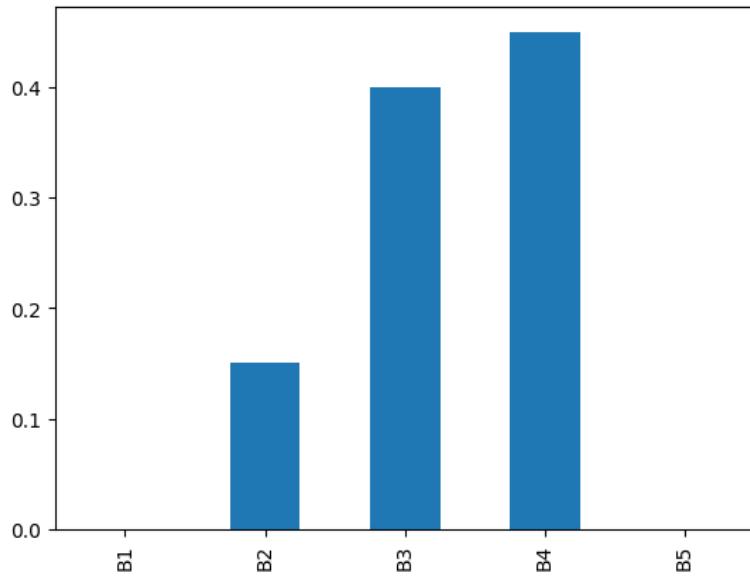


Рис. 50: Биномиальное распределение вероятности двух белых и одного черного шара

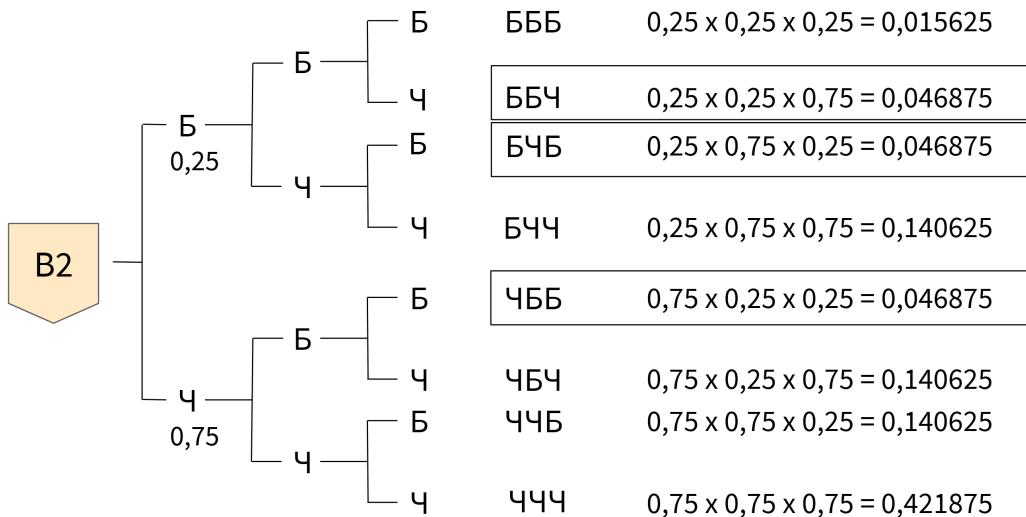


Рис. 51: Биномиальное распределение мешка B2

$$\begin{aligned} P(wwb) &= 0,25 \cdot 0,25 \cdot 0,75 \\ &= 0,046875 \end{aligned}$$

$$\begin{aligned} P(X = 2) &= \binom{3}{2} \cdot 0,25^2 \cdot 0,75^1 \\ &= 3 \cdot 0,046875 = 0,140625 \end{aligned}$$

	prior	likelihood	numerator	posterior		prior	likelihood	numerator	posterior
B1	0.2	0.000000	0.000000	0.00	B1	0.2	0.000000	0.000000	0.00
B2	0.2	0.046875	0.009375	0.15	B2	0.2	0.140625	0.028125	0.15
B3	0.2	0.125000	0.025000	0.40	B3	0.2	0.375000	0.075000	0.40
B4	0.2	0.140625	0.028125	0.45	B4	0.2	0.421875	0.084375	0.45
B5	0.2	0.000000	0.000000	0.00	B5	0.2	0.000000	0.000000	0.00

Рис. 52: Биномиальное распределение мешка B2. Продолжение

такую же апостериорную вероятность, как если бы этот коэффициент не использовался в принципе.

$$\begin{aligned} P(B_2 | wwb) &= \frac{P(wwb | B_2) \cdot P(B_2)}{P(wwb)} = \\ &\frac{P(wwb | B_2) \cdot P(B_2)}{\sum_{i=1}^{n=5} P(wwb | B_n) \cdot P(B_n)} = \\ &\frac{\binom{3}{2} \cdot 0,25^2 \cdot 0,75^1 \cdot 0,20}{\binom{3}{2} (0^2 \cdot 1^1 + 0,25^2 \cdot 0,75^1 + 0,5^2 \cdot 0,5^1 + 0,75^2 \cdot 0,25^1 + 1^2 \cdot 0^1) \cdot 0,20} \end{aligned}$$

Таким образом,

$$P(B_2 | wwb) = \frac{\binom{3}{2} \cdot 0,009375}{\binom{3}{2} \cdot 0,0625} = 0,15$$

Что это нам дает? Мы знаем, что для правдоподобия не обязательно использовать биномиальные коэффициенты и можно ограничиться формулой

$$P(n, k | \theta) \propto \theta^k \cdot (1 - \theta)^{n-k}$$

Запомним этот факт. Позднее он нам пригодится.

4.2 Правдоподобие и вероятность

До сих пор мы отождествляли вероятность и правдоподобие. Однако это не совсем одно и то же. В частности, выше мы сказали, что сумма правдоподобий не равна единице, а с вероятностью так быть не может.

Правдоподобие «привязано» к гипотезам. Дело в том, что правдоподобие «привязано» к гипотезам или параметрам θ и может быть любым, оно не отражает вероятность этой гипотезы относительно других гипотез. Лишь индивидуальное правдоподобие конкретной гипотезы.

Например, ранее мы сказали, что для гипотезы «человек болен», правдоподобие положительного теста составляет 90 процентов, и это значение никак не зависит от другой гипотезы, что «человек здоров».

Аналогичным образом, правдоподобие выпадения двойки на игральной кости зависит исключительно от типа конкретной кости и никак не зависит от других костей.

Функция правдоподобия. На это различие можно взглянуть и так. Когда мы изучаем вероятность случайной величины v , например, биномиальном процессе, то заранее знаем (как бы «фиксируем») параметр p и смотрим, как будет варьироваться вероятность в зависимости от количества испытаний n и количества успехов k . Именно поэтому мы говорим про функцию вероятности (pmf).

Другими словами решаем задачу относительно n и k .

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

В случае правдоподобия, мы «фиксируем» n и k и смотрим, чему может быть равно распределение при различных значениях p (или в нотации формулы Байеса при различных θ).

$$\mathcal{L}(n, k \mid \theta) = \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k}$$

И в этом случае говорят про функцию правдоподобия (likelihood function) \mathcal{L} .

Отношение правдоподобия. Еще раз отметим, что правдоподобие само по себе (без учета априорной вероятности и без учета нормализации) не несет информации о вероятности гипотезы, поскольку не учитывает другие гипотезы и их правдоподобие.

При этом отношение правдоподобия

$$\Lambda_{X|\theta_1} = \frac{\mathcal{L}(X \mid \theta_1)}{\mathcal{L}(X \mid \theta_2)}$$

интерпретировать безусловно можно. Это коэффициент Байеса, который сравнивает правдоподобие двух гипотез.

$$\underbrace{\frac{P(\theta_1 \mid X)}{P(\theta_2 \mid X)}}_{\text{posterior odds}} = \underbrace{\frac{P(\theta_1)}{P(\theta_2)}}_{\text{prior odds}} \cdot \underbrace{\frac{\mathcal{L}(X \mid \theta_1)}{\mathcal{L}(X \mid \theta_2)}}_{\text{Bayes factor}}$$

4.3 100 мешков

Вернемся к задаче с мешками и шарами. В целом мы уже видим, что у нас начинает получаться построить полноценное апостериорное распределение параметра θ и сказать, какие гипотезы являются более вероятными, а какие менее.

Например, мы видим, что с вероятностью 85% два белых и один черный шар появились из мешков три или четыре.

$$P(\theta_3 \cup \theta_4 \mid X) = 0,40 + 0,45 = 0,85$$

При этом конечно хотелось бы (1) оценить распределение с большим количеством гипотез и (2) задать неодинаковую априорную вероятность.

Напишем функции, которые будут сразу создавать таблицу с априорными и апостериорными вероятностями (листинг 38), а также соответствующие графики (листинг 39).

```
def create_table(n_bags, k_whites, n_balls):
    bags = np.arange(1, n_bags + 1)
    white_likelihood = np.linspace(0.0, 1.0, num = n_bags)

    table = pd.DataFrame(index = bags)
    table['prior'] = 1/n_bags
    table['likelihood'] = binom.pmf(k = k_whites, n = n_balls, p =
        white_likelihood)
    table['numerator'] = table['prior'] * table['likelihood']
    prob_data = table['numerator'].sum()
    table['posterior'] = table['numerator'] / prob_data
    return table
```

Листинг 38: Функция для создания таблицы с априорным и апостериорным распределениями

С увеличением количества мешков будет удобно строить линейные графики, а не столбчатые диаграммы.

```
def create_plot(prior, posterior, name):
    prior.plot(label = 'prior')
    posterior.plot(label = 'posterior')
    plt.legend()
    plt.title(name)
    plt.xlabel('Bag')
    plt.ylabel('P(Bag)')
```

Листинг 39: Функция для создания графиков

Один белый шар. Найдем вероятность достать один белый шар из ста мешков (листинг 40 и рисунок 53).

```
bags_100_w = create_table(100, 1, 1)
bags_100_w.head()
```

Листинг 40: Один белый шар из ста мешков. Таблица

	prior	likelihood	numerator	posterior
1	0.01	0.000000	0.000000	0.000000
2	0.01	0.010101	0.000101	0.000202
3	0.01	0.020202	0.000202	0.000404
4	0.01	0.030303	0.000303	0.000606
5	0.01	0.040404	0.000404	0.000808

Рис. 53: Один белый шар из ста мешков. Таблица

Построим график (листинг 41 и рисунок 54).

```
create_plot(bags_100_w['prior'], bags_100_w['posterior'], 'bags_100_w')
```

Листинг 41: Один белый шар из ста мешков. График

Все логично. Так как априорное распределение равномерно, апостериорное распределение при появлении белого шара «подстроилось» под правдоподобие. В первом мешке оно равно нулю, в сотом — максимально.

Площадь под линиями (то есть сумма вероятностей) как в случае априорного, так и в случае апостериорного распределения равна единице.

Два белых шара. Попробуем достать два белых шара (листинг 42).

```
bags_100_ww = create_table(100, 2, 2)
```

Листинг 42: Два белых шара из ста мешков

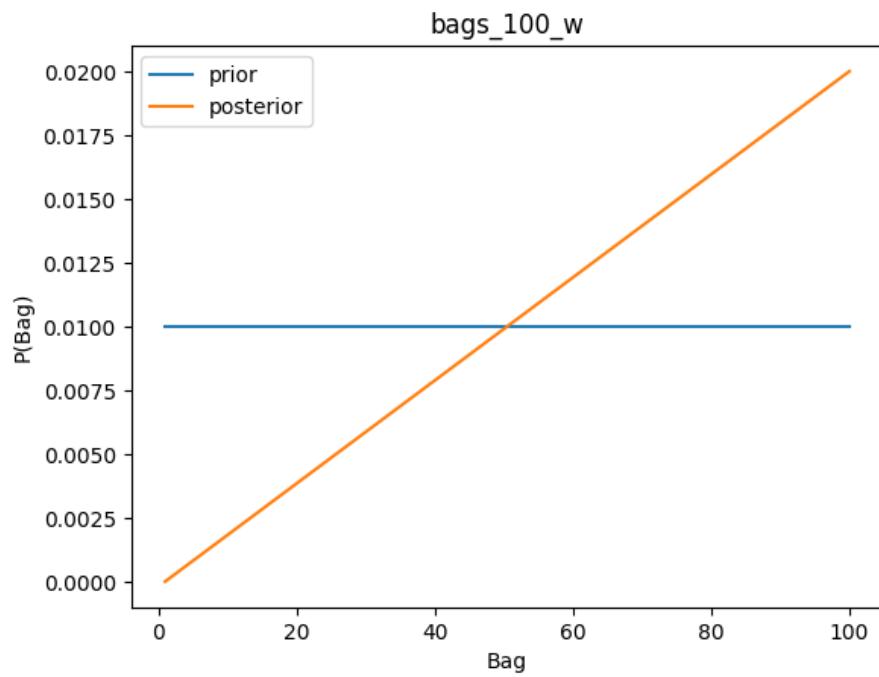


Рис. 54: Один белый шар из ста мешков. График

Построим график (листинг 43 и рисунок 55). При этом априорной вероятностью будет апостериорная вероятность предыдущего испытания, то есть вероятность появления одного белого шара.

```
create_plot(bags_100_w['posterior'], bags_100_ww['posterior'],  
           bags_100_ww)
```

Листинг 43: Два белых шара из ста мешков. График

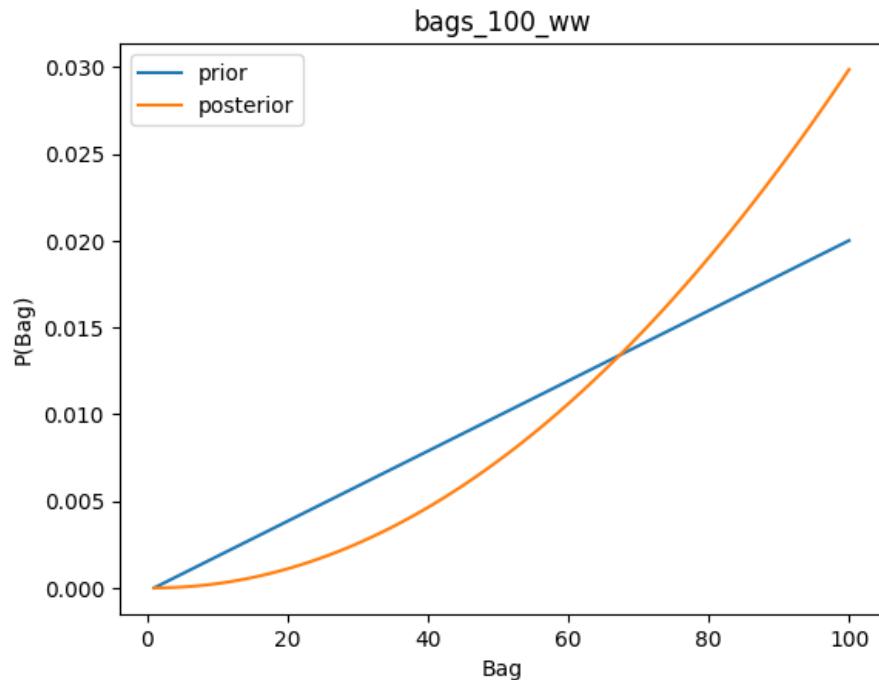


Рис. 55: Два белых шара из ста мешков. График

Апостериорные вероятности стали «тяготеть» к мешкам со старшими индексами.

Два белых и один черный шар. Достанем еще один черный шар (листинг 44 и рисунок 56).

```
bags_100_wwb = create_table(100, 2, 3)
create_plot(bags_100_wwb['posterior'], bags_100_wwb['posterior'], ,
           bags_100_wwb)
```

Листинг 44: Два белых и один черный шар из ста мешков. График

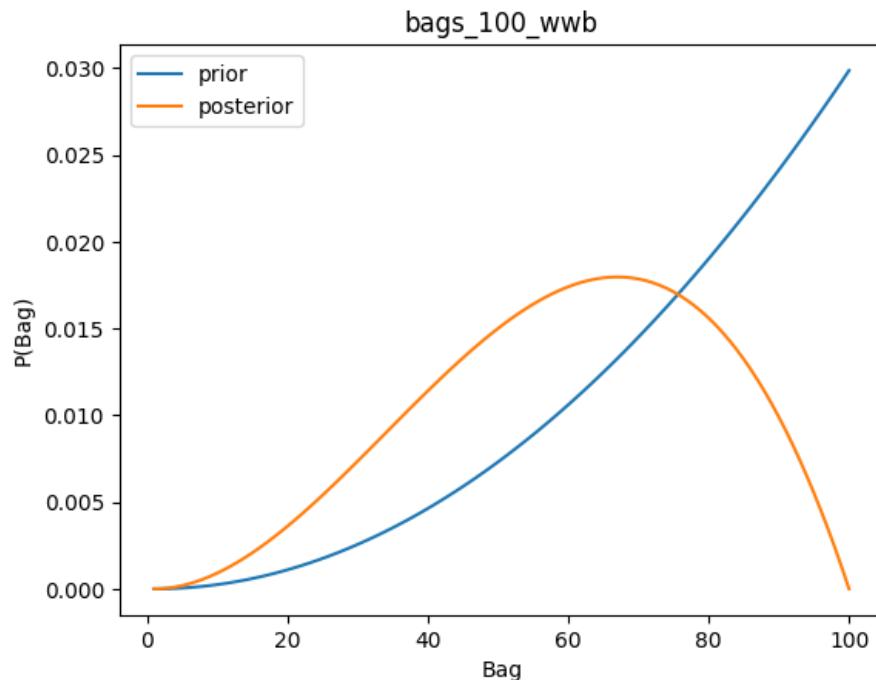


Рис. 56: Два белых и один черный шар из ста мешков. График

На графике выше видно, что с учетом новых данных 100-й мешок также невозможен, для него правдоподобие белого шара равно единице, а мы только что достали черный шар.

Оценим апостериорный максимум и максимальное правдоподобие с помощью метода `.idxmax()`, который выведет индекс мешка с максимальным значением (листинг 45).

```
MAP = bags_100_wwb['posterior'].idxmax()
MAP

Output: 67

MLE = bags_100_wwb['likelihood'].idxmax()
MLE

Output: 67
```

Листинг 45: MAP и MLE-оценка двух белых и одного черного шара

Как мы уже говорили, эти оценки совпадают, так как априорное распределение равномерно. При этом численно они разумеются не одинаковы (листинг 46 и рисунок 57).

```
bags_100_wwb.iloc[65:68, [1,3]]
```

Листинг 46: Численная МАР и МЛЕ-оценка

	likelihood	posterior
66	0.444141	0.017947
67	0.444444	0.017959
68	0.444135	0.017947

Рис. 57: Численная МАР и МЛЕ-оценка

Максимальное правдоподобие биномиального распределения. Полученную МЛЕ-оценку можно выразить как отношение к количеству гипотез, т.е. $67/100 = 0,67$ и это как раз то значение p или θ , относительно которого мы находим максимум функции правдоподобия (при фиксированных данных, то есть n и k).

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(n, k | \theta) = \binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} = \binom{3}{2} \cdot \theta^2 \cdot (1 - \theta)^{3-2}$$

Максимальное правдоподобие биномиальной функции можно найти аналитически.

$$\theta_{MLE} = \frac{k}{n} = \frac{2}{3} \approx 0,67$$

Приведем несложное доказательство. Возьмем логарифм правдоподобия. Такая функция называется логарифмической функцией правдоподобия (log-likelihood, ℓ).

$$\begin{aligned} \ell(\theta) &= \log \left(\binom{n}{k} \cdot \theta^k \cdot (1 - \theta)^{n-k} \right) \\ &= \log \binom{n}{k} + \log \theta^k + \log(1 - \theta)^{n-k} \\ &= \log \binom{n}{k} + k \log \theta + (n - k) \log(1 - \theta) \end{aligned}$$

Найдем производную относительно θ . Вспомним, что первое слагаемое — это константа, а производная натурального логарифма $(\ln x)' = \frac{1}{x}$. Тогда

$$\frac{d\ell}{d\theta} = \frac{k}{\theta} - \frac{n - k}{1 - \theta}$$

Теперь приравняем производную к нулю и найдем максимум функции правдоподобия θ_{MLE} .

$$\begin{aligned} \frac{d\ell}{d\theta_{MLE}} &= 0 \\ \frac{k}{\theta_{MLE}} - \frac{n - k}{1 - \theta_{MLE}} &= 0 \\ \frac{n - k}{1 - \theta_{MLE}} &= \frac{k}{\theta_{MLE}} \\ (n - k)\theta_{MLE} &= k(1 - \theta_{MLE}) \\ n\theta_{MLE} &= k \\ \theta_{MLE} &= \frac{k}{n} \end{aligned}$$

Логарифмическая функция правдоподобия. Почему мы так легко взяли логарифм от функции правдоподобия? Дело в том, что функция логарифма монотонно возрастает (листинг 47 и рисунок 58).

```
x = np.linspace(0.001, 15)
y = np.log(x)
plt.plot(x,y);
```

Листинг 47: Функция логарифма

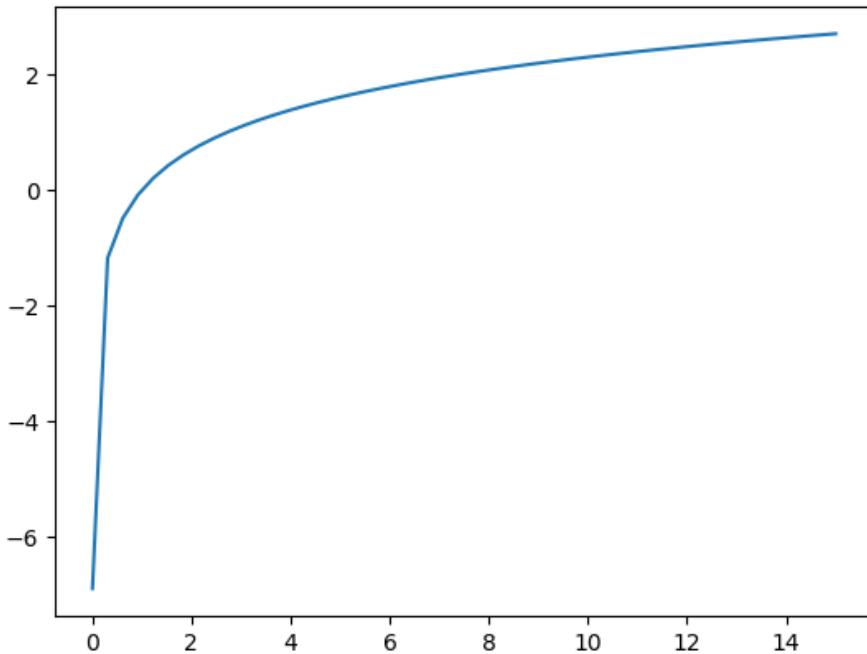


Рис. 58: Функция логарифма

Как следствие (листинг 48 и рисунок 59), наибольшее значение функции правдоподобия $\mathcal{L}(X | \theta)$ всегда будет одновременно наибольшим значением $\ell(X | \theta)$.

```

L = bags_100_wwb['likelihood']
l = np.log(bags_100_wwb['likelihood']) + 10 ** (-5)

fig, ax1 = plt.subplots()
color = 'tab:blue'
ax1.set_xlabel('theta')
ax1.set_ylabel('likelihood, L', color = color)
ax1.plot(L, color=color)
ax1.tick_params(axis = 'y', labelcolor = color)

ax2 = ax1.twinx()

color = 'tab:orange'
ax2.set_ylabel('log-likelihood, l', color = color)
ax2.plot(l, color=color)
ax2.tick_params(axis = 'y', labelcolor = color)

fig.tight_layout()
plt.show()

```

Листинг 48: Логарифмическая функция правдоподобия

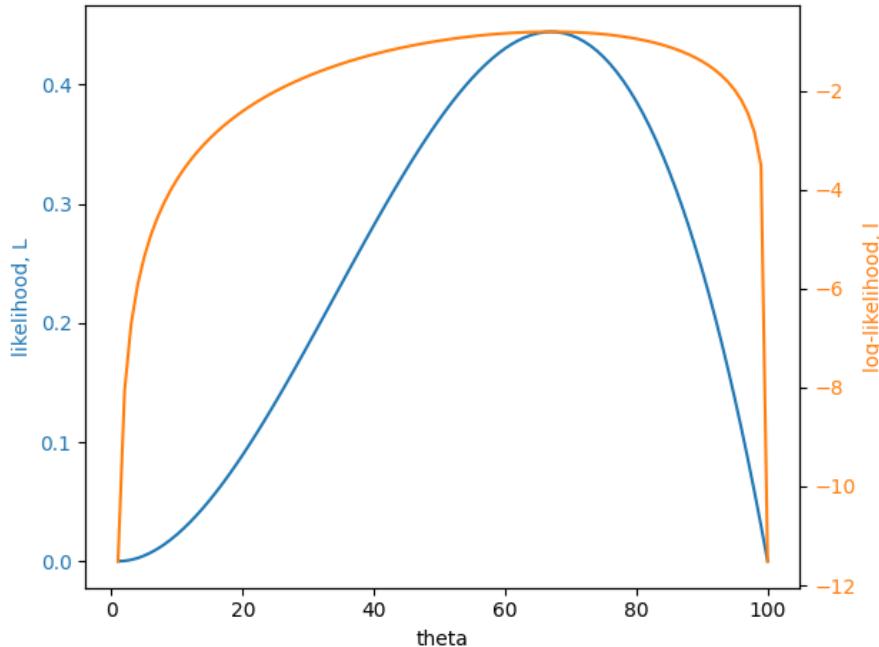


Рис. 59: Логарифмическая функция правдоподобия

Семь белых шаров из десяти. Достанем семь белых шаров из десяти (листинг 49 и рисунок 60).

```

bags_7w_10 = create_table(100, 7, 10)
create_plot(bags_7w_10['prior'], bags_7w_10['posterior'], 'bags_7w_10')

```

Листинг 49: Семь белых шаров из десяти

Ожидаемо, наиболее вероятным будет 70-ый мешок (листинг 50).

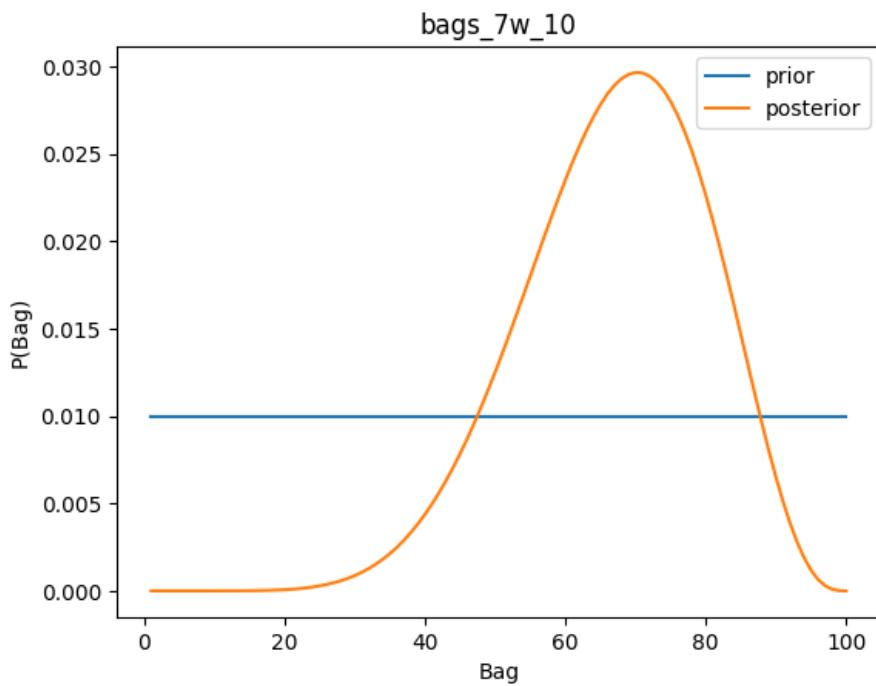


Рис. 60: Семь белых шаров из десяти

```
MLE = bags_7w_10['likelihood'].idxmax()
MLE

Output: 70
```

Листинг 50: Оценка максимального правдоподобия семи белых шаров из десяти

Это соответствует аналитической MLE-оценке биномиального распределения

$$\theta_{MLE} = \frac{k}{n} = \frac{7}{10} = 0,7$$

Другая априорная вероятность. Теперь зададим другую априорную вероятность. Для простоты пусть это будет функция, которая равномерно возрастает до 50-го мешка, а затем также равномерно убывает. Назовем ее треугольником (листинг 51).

```
up = np.arange(50)
down = np.arange(50, 0, -1)
up_and_down = np.append(up, down)
triangle_prior = pd.Series(up_and_down)
triangle_prior /= triangle_prior.sum()
```

Листинг 51: Другая априорная вероятность

Создадим уже знакомую таблицу (листинг 52).

```

n_bags = 100

bags = np.arange(1, n_bags + 1)
white_likelihood = np.linspace(0, 1, num = n_bags)

bags_triangle = pd.DataFrame(index = bags)

bags_triangle['prior'] = triangle_prior
bags_triangle['likelihood'] = binom.pmf(k = 7, n = 10, p =
    white_likelihood)

bags_triangle['numerator'] = bags_triangle['prior'] * bags_triangle['
    likelihood']
prob_data = bags_triangle['numerator'].sum()
bags_triangle['posterior'] = bags_triangle['numerator'] / prob_data

```

Листинг 52: Другая априорная вероятность. Таблица

Построим график (листинг 53 и рисунок 61).

```

create_plot(bags_triangle['prior'], bags_triangle['posterior'], '
    bags_triangle')

```

Листинг 53: Другая априорная вероятность. График

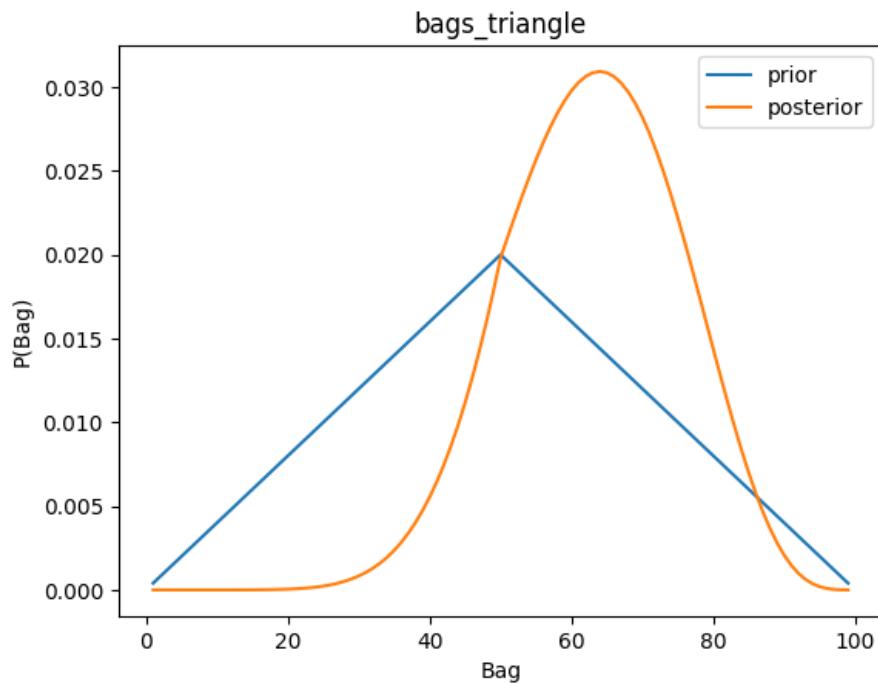


Рис. 61: Другая априорная вероятность. График

Сравним априорный максимум (листинг 54), правдоподобие (листинг 55) и апостериорный максимум (листинг 56).

```
bags_triangle['prior'].idxmax()
```

```
Output: 50
```

Листинг 54: Другая априорная вероятность. Априорный максимум

```
bags_triangle['likelihood'].idxmax()
```

```
Output: 70
```

Листинг 55: Другая априорная вероятность. Правдоподобие

```
bags_triangle['posterior'].idxmax()
```

```
Output: 64
```

Листинг 56: Другая априорная вероятность. Апостериорный максимум

Поясним результаты. Изначально мы считали 50-й мешок наиболее вероятным, при этом исходя из данных наиболее правдоподобными был 70-й мешок. Как следствие, апостериорный максимум сместился в сторону наиболее правдоподобного мешка, однако после появления трех шаров не совпал с ним.

4.4 Бесконечное количество гипотез

Итак мы перешли от оценки отдельных гипотез к оценке их (дискретного) распределения. Продолжим использовать пример с мешками и шарами. Только теперь попробуем перейти от конечного числа мешков или гипотез к бесконечным θ на интервале от 0 до 1. Шаров при этом по-прежнему два: белый и черный.

Примечание. Интервал возможных значений гипотез от 0 до 1 удобен тем, что, например, мы смогли бы ответить на вопрос, с какой вероятностью параметр θ примет значение $\theta \geq 0,5$, то есть найти *вероятность вероятности*.

Априорная вероятность. Вначале займемся априорной вероятностью гипотезы θ . Если в дискретном случае мы говорили про $p(\theta)$, то теперь можем задать функцию плотности априорной вероятности (probability density function, pdf) $f(\theta)$.

Пусть в нашем новом примере такая функция задана следующим образом.

$$f(\theta) = 2\theta$$

Построим график (листинг 57 и рисунок 62).

```

def prior(theta):
    return 2 * theta

theta = np.linspace(0, 1, 100)
plt.plot(theta, prior(theta), label = r'$f(\theta) = 2\theta$')
plt.legend()
plt.title('Prior')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')
plt.grid()
plt.show()

```

Листинг 57: Бесконечное количество гипотез

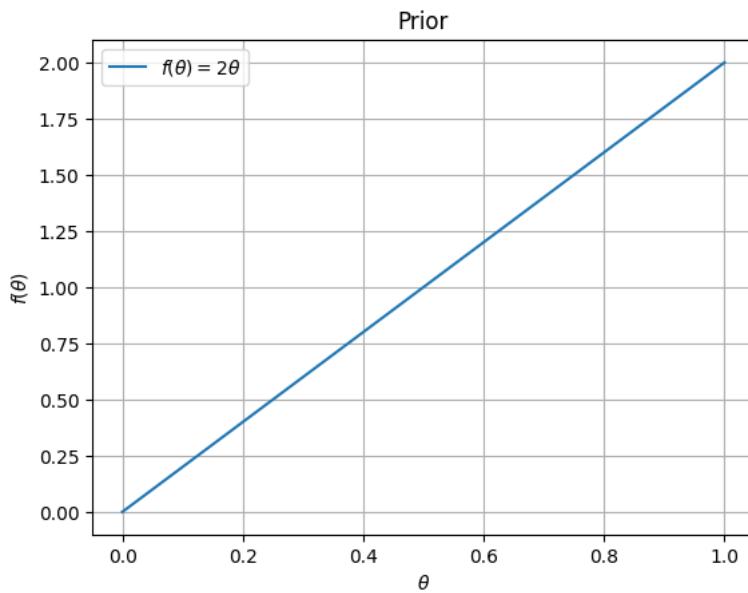


Рис. 62: Бесконечное количество гипотез

Другими словами, мы считаем, что изначально более вероятными являются приближающиеся к единице гипотезы. Убедимся в этом, вычислив, как мы сказали выше, вероятность того, что параметр θ примет значение $\theta \geq 0,5$.

Для этого нам нужно найти интегральную сумму отрезков длиной $d\theta$ на интервале $0,5 \leq \theta \leq 1$.

$$\int_{0,5}^1 2\theta d\theta = \frac{3}{4} = 0,75$$

Посмотрим на график (листинг 58 и рисунок 63).

```

fill_theta = np.linspace(0.5, 1, 100)
plt.plot(theta, prior(theta), label = r'$f(\theta) = 2\theta$')
plt.fill_between(fill_theta, prior(fill_theta), alpha = 0.3)
plt.legend()
plt.title(r'Prior, $P(\theta) \geq 0{,}5$')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')
plt.grid()
plt.show()

```

Листинг 58: Вероятность параметра $\theta \geq 0,5$

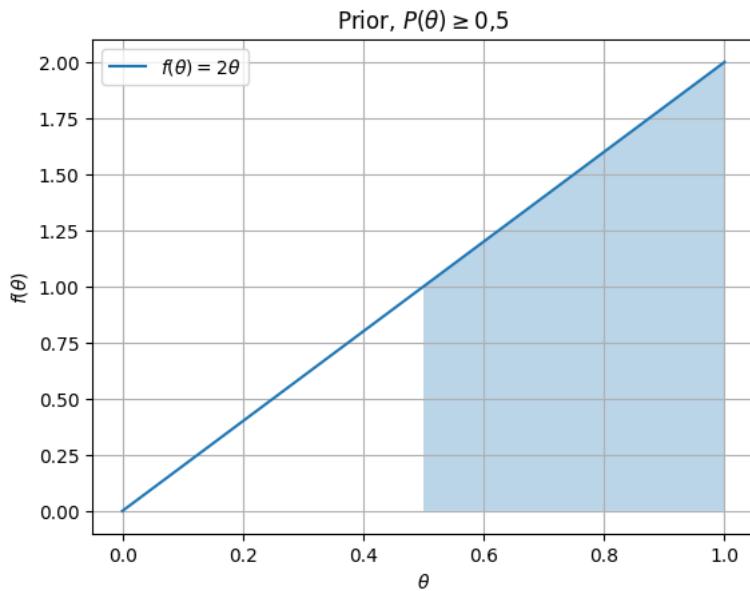


Рис. 63: Вероятность параметра $\theta \geq 0,5$

Вычислим интеграл с помощью Питона (листинг 59).

```

from scipy.integrate import quad
quad(prior, 0.5, 1)[0]

Output: 0.75

```

Листинг 59: Интеграл параметра $\theta \geq 0,5$. Априорная вероятность

Правдоподобие. Правдоподобие, как и априорная вероятность, может быть бесконечным, однако так как мы решили, что по-прежнему достаем два вида шаров из каждого из мешков, то правдоподобие дискретно и следует биномиальному процессу.

$$p(X | \theta) \propto \theta^k (1 - \theta)^{n-k}$$

Один белый шар. Предположим, что мы достали один белый шар. В этом случае числитель формулы Байеса будет равен

$$f(\theta) \cdot p(w|\theta) = 2\theta \cdot \theta^1 \cdot (1 - \theta)^{1-1} = 2\theta^2$$

Формула полной вероятности. Самое интересное же произойдет в знаменателе или формуле полной вероятности, где нам нужно будет найти интеграл всех возможных значений θ при появлении одного белого шара.

$$p(w) = \int_0^1 2\theta d\theta \cdot \theta = \int_0^1 2\theta^2 d\theta = \frac{2}{3}$$

Апостериорная вероятность. Таким образом, формула Байеса для бесконечного числа гипотез и дискретного правдоподобия будет равна

$$f(\theta | x)d\theta = \frac{p(x | \theta) \cdot f(\theta)}{\int_a^b p(x | \theta)f(\theta)d\theta} = \frac{p(x | \theta)f(\theta)}{p(x)}$$

В нашем случае,

$$f(\theta | w) = \frac{2\theta^2}{\int_0^1 2\theta^2 d\theta} = \frac{2\theta^2}{\frac{2}{3}} = 3\theta^2$$

Графически, ожидаемо, апостериорная вероятность еще больше сместилась в сторону гипотез, близких к единице (листинг 60 и рисунок 64).

```
def posterior(theta):
    return 3 * theta**2

plt.plot(theta, posterior(theta), label = r'$f(\theta) = 3\theta^2$')
plt.fill_between(fill_theta, posterior(fill_theta), alpha = 0.3)
plt.legend()
plt.title(r'Posterior, $P(\theta) \geq 0{,}5$')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')
plt.grid()
plt.show()
```

Листинг 60: Апостериорная вероятность параметра $\theta \geq 0,5$

Убедимся в этом. Найдем вероятность того, что $\theta \geq 0,5$.

$$\int_{0,5}^1 3\theta^2 d\theta = \frac{7}{8} = 0,875$$

Вычислим интеграл с помощью Питона (листинг 61).

```
quad(posterior, 0.5, 1)[0]

Output: 0.875
```

Листинг 61: Интеграл параметра $\theta \geq 0,5$. Апостериорная вероятность

При этом апостериорный максимум будет равен $\theta_{MAP} = 1$.

Равномерная априорная вероятность. Теперь рассмотрим более простой случай равномерной априорной вероятности. Такая вероятность задана функцией

$$f(\theta) = 1$$

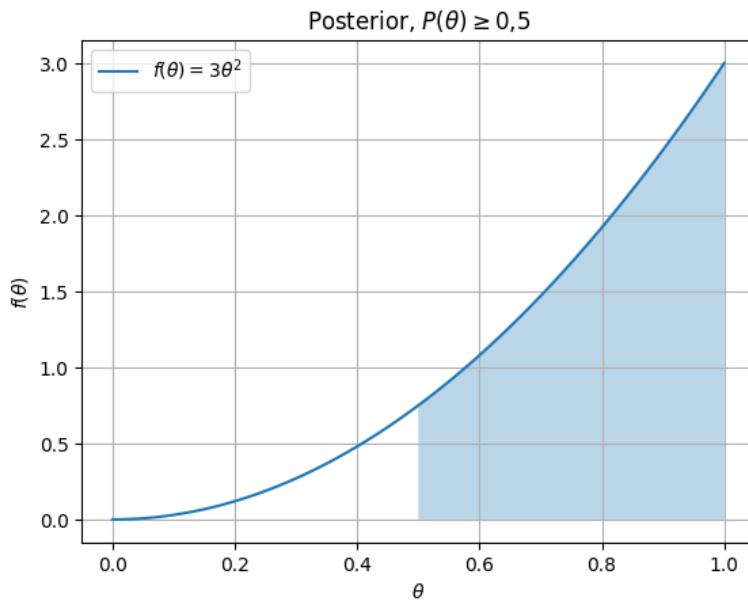


Рис. 64: Апостериорная вероятность параметра $\theta \geq 0,5$

Построим график (листинг 62 и рисунок 65).

```
from scipy.stats import uniform
plt.plot(theta, uniform.pdf(theta), label = r'$f(\theta) = 1$')
plt.legend()
plt.title('Uniform prior')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')
plt.grid()
plt.show()
```

Листинг 62: Равномерная априорная вероятность

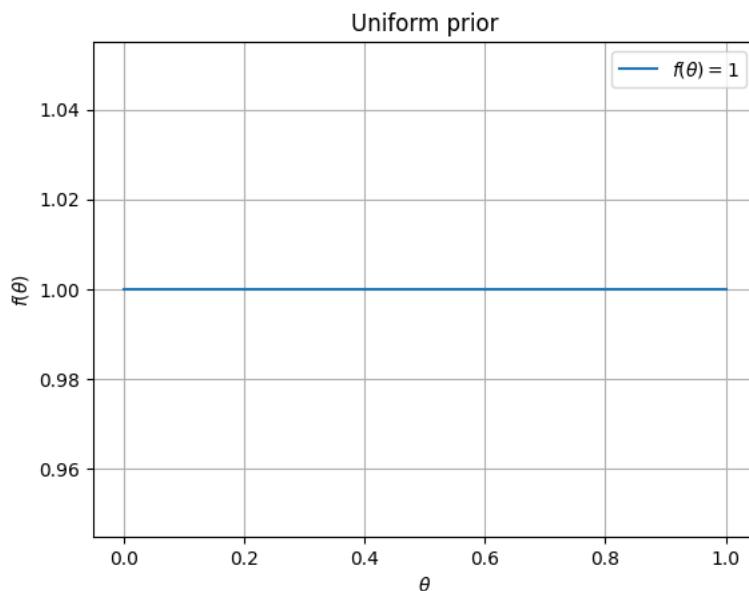


Рис. 65: Равномерная априорная вероятность

Найдем функцию апостериорной вероятности появления одного белого шара.

$$f(\theta | w) = \frac{1 \cdot \theta}{\int_0^1 \theta d\theta} = \frac{\theta}{\frac{1}{2}} = 2\theta$$

Теперь, как и в предыдущих примерах, найдем функцию вероятности появления двух белых и одного черного шара. В этом случае правдоподобие будет равно

$$P(wwb | \theta) = \theta^2 \cdot (1 - \theta)^{3-1},$$

а функция апостериорной вероятности примет вид

$$f(\theta | wwb) = \frac{1 \cdot (\theta^2(1 - \theta))}{\int_0^1 \theta^2(1 - \theta) d\theta} = \frac{\theta^2(1 - \theta)}{\frac{1}{12}} = 12\theta^2(1 - \theta)$$

Посмотрим на график (листинг 63 и рисунок 66).

```
def posterior(theta):
    return 12 * theta**2 * (1 - theta)

plt.plot(theta, posterior(theta), label = r'$f(\theta) = 12\theta^2(1-\theta)$')
plt.fill_between(fill_theta, posterior(fill_theta), alpha = 0.3)
plt.vlines(x = 2/3, ymin = 0, ymax = 1.78, colors = 'r', label = r'$\theta_{MAP}$')
plt.legend()
plt.title(r'Posterior, $P(\theta) \geq 0{,}5$')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')
plt.grid()
plt.show()
```

Листинг 63: Функция апостериорной вероятности $12\theta^2(1 - \theta)$ и апостериорный максимум

Снова найдем вероятность $\theta \geq 0,5$ (листинг 64).

$$\int_{0,5}^1 12\theta^2(1 - \theta) d\theta = \frac{11}{16} = 0,6875$$

```
quad(posterior, 0.5, 1)[0]
```

```
Output: 0.6875
```

Листинг 64: Интеграл параметра $\theta \geq 0,5$ при равномерной априорной вероятности

Апостериорный максимум. Найдем апостериорный максимум θ_{MAP} . Возьмем производную.

$$f''(\theta) = (12\theta^2(1 - \theta))' = -12\theta(3\theta - 2)$$

Приравняв производную к нулю, найдем критические точки $\theta_{1,2} = 0, \frac{2}{3}$. Так как на интервале $(0; \frac{2}{3})$ значение производной положительно (функция возрастает), а на интервалах $(-\infty; 0)$ и $(\frac{2}{3}; \infty)$ отрицательно (функция убывает), то $\theta_{MAP} = \frac{2}{3}$, что соответствует приведенному графику.

Обратим внимание, что это ожидаемо соответствует максимальному значению биномиального распределения при $k = 2, n = 3$.

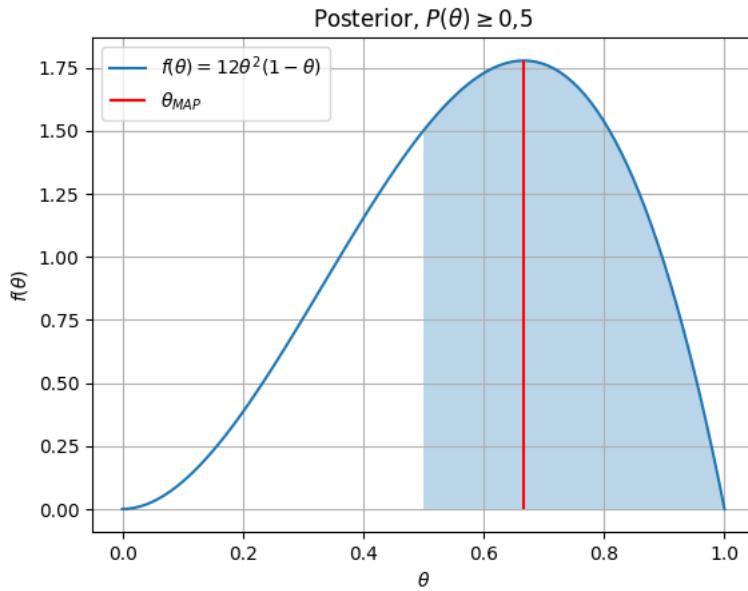


Рис. 66: Функция апостериорной вероятности $12\theta^2(1 - \theta)$ и апостериорный максимум

5 Биномиальное и бета-распределение

Теперь посмотрим, как можно построить непрерывное апостериорное распределение и при этом избежать вычисления интеграла.

5.1 Бета-распределение

Рассмотрим бета-распределение (beta distribution) с вещественными положительными параметрами α и β .

$$X \sim Beta(\alpha, \beta)$$

Функция плотности. Функция плотности бета-распределения имеет вид

$$f(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}$$

Знаменатель распределения представляет собой нормализующую константу и называется бета-функцией (beta function, B), которая и дала название самому распределению.

$$\begin{aligned} f(\theta; \alpha, \beta) &= const \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

Одновременно бета-функция может быть определена через гамма-функцию Γ с теми же параметрами.

$$f(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Гамма-функция. Гамма-функция (gamma function, Γ) является обобщением понятия факториала на множества действительных и комплексных значений, и для нее справедливо, что

$$\Gamma(z) = (z - 1)!$$

или, по свойству $\Gamma(z + 1) = z\Gamma(z)$,

$$\Gamma(z + 1) = z\Gamma(z) = z(z - 1)! = z!$$

Семейство распределений. Распределения характеризуются параметрами, которые влияют на вид или форму этого распределения.

Например, в биномиальном распределении вероятность успеха, параметр p , влияет на матожидание, коэффициент асимметрии и другие свойства распределения. В нормальном распределении на него влияют матожидание μ и дисперсия σ^2 . В бета-распределении — параметры α и β .

Совокупность распределений со всеми возможными значениями данных параметров называется **семейством распределений** (distribution family).

Более формально речь идет о множестве распределений \mathcal{P} , в которых параметр θ отдельного распределения \mathcal{P}_θ принадлежит множеству $\Theta \subseteq \mathbb{R}^s$.

$$\mathcal{P} = \{\mathcal{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^s\}$$

Таким образом, можно говорить о семействе биномиальных, нормальных, бета- и других распределений.

Особенности распределения. В частности, бета-распределение представляет собой семейство распределений, ограниченное на промежутке $\theta \in [0, 1]$ (**носитель** (support) распределения) и в зависимости от значений параметров α и β может быть достаточно разнообразным.

Приведем несколько примеров (листинг 65 и рисунок 67).

```
plt.plot(theta, beta.pdf(theta, 0.2, 0.4), label = r'$\alpha = 0{,}2$, \
    beta = 0{,}4$')
plt.plot(theta, beta.pdf(theta, 1, 1), label = r'$\alpha = 1, \beta = 1$')
plt.plot(theta, beta.pdf(theta, 2, 5), label = r'$\alpha = 5, \beta = 2$')
plt.plot(theta, beta.pdf(theta, 21, 7), label = r'$\alpha = 21, \beta = 7$')
plt.legend()
plt.title(r'Beta distribution family')
plt.xlabel(r'$\theta$')
plt.ylabel('PDF')
plt.grid()
plt.show()
```

Листинг 65: Семейство бета-распределений

Обратим внимание на три момента:

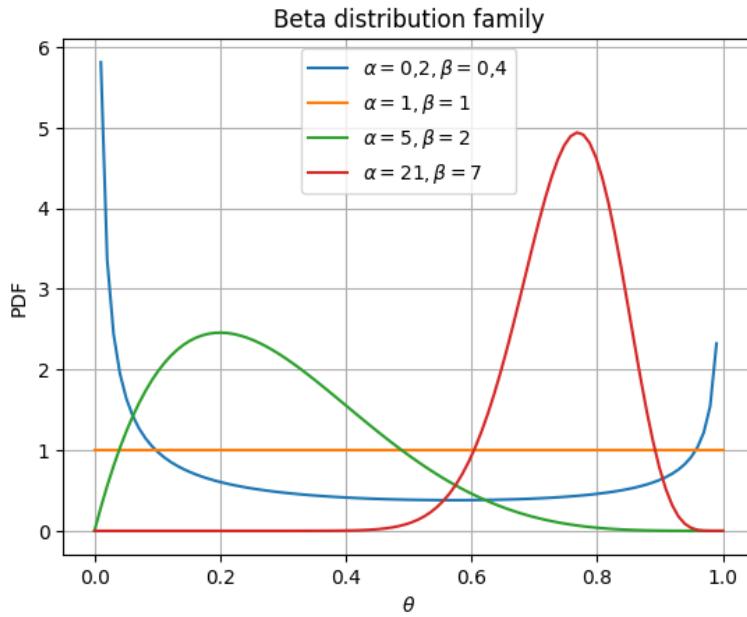


Рис. 67: Семейство бета-распределений

- бета-распределение с параметрами $Beta(1, 1)$ идентично равномерному распределению $U(0, 1)$ (оранжевая кривая); кроме того,
- при увеличении β основная плотность вероятности смещается к нулю, при увеличении α — к единице (например, зеленая кривая); и что особенно важно
- с ростом α и β «уверенность» в определенном значении θ (высота пика) возрастает (сравните зеленую и красную кривые).

Также отметим, что бета-распределение не может быть бимодальным (иметь более одного пика).

Вывод напрашивается сам собой. Бета-распределение удобно использовать в качестве априорного распределения в связке с биномиальным правдоподобием, а параметры α и β считать «успехами» или «неудачами» до того, как мы начали собирать данные.

5.2 Бета- и биномиальное сопряженные распределения

Понятие сопряженности. Более того, если в формуле Байеса «соединить» априорное бета-распределение и биномиальное правдоподобие, то апостериорным будет распределение из семейства бета-распределений, но уже с новыми параметрами α и β .

Другими словами,

- если правдоподобие распределено как $\mathcal{L}(n, k \mid \theta) \sim Binom(n, \theta)$;
- априорное распределение как $P(\theta \mid \alpha, \beta) \sim Beta(\alpha, \beta)$; то
- апостериорным распределением будет $P(\theta \mid k, n, \alpha, \beta) \sim Beta(\alpha + k, \beta + n - k)$.

По сути мы прибавляем априорные успехи к успехам в полученных данных, а априорные неудачи соответственно к неудачам.

Такое свойство, когда априорное и апостериорное распределения принадлежат к одному семейству, называется **сопряженностью** (conjugacy) распределения и правдоподобия.

При этом бета-распределение и биномиальное распределение также называют **парой сопряженных распределений** (conjugate pair).

Вернемся к примеру с появлением двух белых и одного черного шара, и равномерным априорным распределением (листинг 66 и рисунок 68).

```
a, b = 1, 1
k, n = 2, 3

prior = beta.pdf(theta, a, b)
posterior = beta.pdf(theta, a+k, b+n-k)

plt.plot(theta, prior, label = 'prior')
plt.plot(theta, posterior, label = 'posterior')

plt.vlines(x = posterior.argmax()/len(theta), ymin = 0, ymax = 1.78,
           colors = 'r', label = r'\theta_{MAP}')

plt.title('Beta prior and posterior distributions, wwb')
plt.xlabel(r'\theta')
plt.ylabel(r'f(\theta)')

plt.legend()
plt.show()
```

Листинг 66: Априорное бета-распределение и биномиальное правдоподобие

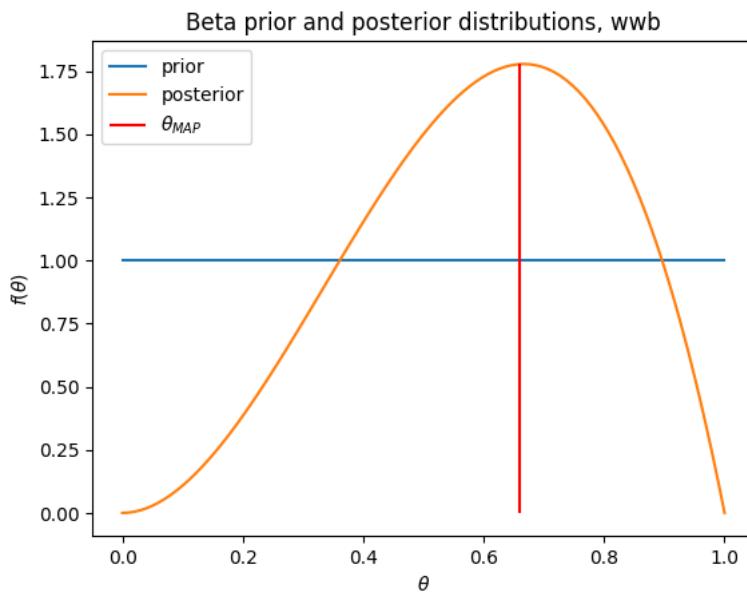


Рис. 68: Априорное бета-распределение и биномиальное правдоподобие

Вновь найдем вероятность $\theta \geq 0,5$, но уже используя свойство сопряженности.

```
1 - beta.cdf(0.5, a+k, b+n-k)
Output: 0.6875
```

Листинг 67: $BetaBin(6, 1, 1)$

Как мы видим, результаты (в частности, апостериорный максимум и вероятность $\theta \geq 0,5$) идентичны предыдущим вычислениям «вручную».

Посмотрим, почему это работает. Приведем вначале интуитивное, а затем более строгое доказательство сопряженности бета- и биномиального распределений.

Интуитивное доказательство. Вновь выпишем априорное бета-распределение

$$Beta(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Так как первый множитель — это константа, то

$$\begin{aligned} Beta(\alpha, \beta) &= const \cdot \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \end{aligned}$$

и биномиальное правдоподобие

$$P(X | \theta) \propto \theta^k (1-\theta)^{n-k}$$

Вспомним, что формулу Байеса можно записать как

$$P(\theta | X) = \frac{P(X | \theta) \cdot P(\theta)}{P(X)} \propto P(X | \theta) \cdot P(\theta)$$

Подставим соответствующие априорное распределение и правдоподобие

$$\begin{aligned} P(\theta | X) &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \cdot \theta^k (1-\theta)^{n-k} \\ &\propto \theta^{\alpha+k-1} (1-\theta)^{n+\beta-k-1} \end{aligned}$$

Если положить $\alpha' = \alpha + k$ и $\beta' = n + \beta - k$, то получим

$$Beta(\alpha', \beta') = \frac{\theta^{\alpha'-1} (1-\theta)^{\beta'-1}}{B(\alpha', \beta')},$$

что очевидно также является бета-распределением.

Более строгое доказательство. Приведем полное доказательство.

Числитель формулы Байеса. Вспомним, что числитель формулы Байеса есть совместная вероятность данных X и параметра θ . Тогда, для непрерывной случайной величины, знаменатель будет интегралом совместных вероятностей всех возможных значений параметра и полученных данных.

$$P(\theta | X) = \frac{P(X | \theta) \cdot P(\theta)}{P(X)} = \frac{P(X, \theta)}{\int_0^1 P(X, \theta)}$$

Биномиальное правдоподобие будет равно

$$P(X | \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} = \frac{n!}{(n - k)! k!} \cdot \theta^k (1 - \theta)^{n-k}$$

Поскольку $\Gamma(z + 1) = z!$, то

$$\binom{n}{k} = \frac{n!}{(n - k)! k!} = \frac{\Gamma(n + 1)}{\Gamma(n - k + 1)\Gamma(k + 1)}$$

Априорное распределение, выраженное через гамма-функцию, будет иметь вид

$$P(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

В этом случае совместную вероятность данных X и параметра θ (числитель формулы Байеса) можно выразить как

$$P(X, \theta) = \frac{\Gamma(n + 1)}{\Gamma(n - k + 1)\Gamma(k + 1)} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha+k-1} (1 - \theta)^{n+\beta-k-1}$$

Знаменатель формулы Байеса. Найдем вероятность данных $P(X)$ (знаменатель). Обозначим первые два множителя через γ .

$$\gamma = \frac{\Gamma(n + 1)}{\Gamma(n - k + 1)\Gamma(k + 1)} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Тогда,

$$P(X, \theta) = \gamma \cdot \theta^{\alpha+k-1} (1 - \theta)^{n+\beta-k-1}$$

Положим $\alpha' = \alpha + k$ и $\beta' = n + \beta - k$

$$P(X, \theta) = \gamma \cdot \theta^{\alpha'-1} (1 - \theta)^{\beta'-1}$$

Умножим это выражение на $\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}$ так, чтобы

$$P(X, \theta) = \frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha')\Gamma(\beta')} \cdot \theta^{\alpha'-1} (1 - \theta)^{\beta'-1} \cdot \gamma \cdot \frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha' + \beta')}$$

При этом вспомним, что мы хотим найти вероятность данных $P(X)$, то есть интеграл всех возможных значений θ . Другими словами,

$$\begin{aligned} P(X) &= \int_0^1 P(X, \theta) d\theta \\ &= \int_0^1 \left(\frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha')\Gamma(\beta')} \cdot \theta^{\alpha'-1} (1 - \theta)^{\beta'-1} \cdot \gamma \cdot \frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha' + \beta')} \right) d\theta \end{aligned}$$

Вторые два множителя при интегрировании относительно θ являются константой.

$$\gamma \cdot \frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha' + \beta')} \cdot \int_0^1 \left(\frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha')\Gamma(\beta')} \cdot \theta^{\alpha'-1} (1 - \theta)^{\beta'-1} \right) d\theta$$

Одновременно первые два множителя — есть интеграл бета-распределения $Beta(\alpha', \beta')$, который по определению равен единице.

$$\int_0^1 \left(\frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha')\Gamma(\beta')} \cdot \theta^{\alpha'-1} (1 - \theta)^{\beta'-1} \right) d\theta = 1$$

Тогда,

$$P(X) = \gamma \cdot \frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha' + \beta')}$$

Апостериорная вероятность. Подставим результаты в формулу апостериорной вероятности.

$$\begin{aligned} P(\theta | X) &= \frac{\gamma \cdot \theta^{\alpha'-1} (1-\theta)^{\beta'-1}}{\gamma \cdot \frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha'+\beta')}} \\ &= \frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha')\Gamma(\beta')} \cdot \theta^{\alpha'-1} (1-\theta)^{\beta'-1} \\ &= \frac{\theta^{\alpha'-1} (1-\theta)^{\beta'-1}}{B(\alpha', \beta')} = Beta(\alpha', \beta') \end{aligned}$$

5.3 Бета-биномиальное распределение

Сопряженность бета- и биномиального распределений можно продемонстрировать иначе.

Числитель формулы Байеса. Вернем выражение, которое мы обозначили через γ .

$$P(X, \theta) = \frac{\Gamma(n+1)}{\Gamma(n-k+1)\Gamma(k+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha+k-1} (1-\theta)^{n+\beta-k-1}$$

Вспомним, что

$$P(X, \theta) = \underbrace{\frac{\Gamma(n+1)}{\Gamma(n-k+1)\Gamma(k+1)}}_{\binom{n}{k}} \cdot \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\frac{1}{B(\alpha, \beta)}} \cdot \theta^{\alpha+k-1} (1-\theta)^{n+\beta-k-1}$$

Тогда,

$$P(X, \theta) = \binom{n}{k} \frac{\theta^{\alpha+k-1} (1-\theta)^{n+\beta-k-1}}{B(\alpha, \beta)}$$

Знаменатель формулы Байеса. Проделаем аналогичную работу со знаменателем.

$$\begin{aligned} P(X) &= \gamma \cdot \frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha'+\beta')} \\ P(X) &= \underbrace{\frac{\Gamma(n+1)}{\Gamma(n-k+1)\Gamma(k+1)}}_{\binom{n}{k}} \cdot \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\frac{1}{B(\alpha, \beta)}} \cdot \underbrace{\frac{\Gamma(\alpha')\Gamma(\beta')}{\Gamma(\alpha'+\beta')}}_{B(\alpha', \beta')} \end{aligned}$$

Таким образом, если вернуть $\alpha' = \alpha + k$ и $\beta' = n + \beta - k$,

$$P(X) = \binom{n}{k} \frac{B(\alpha+k, n+\beta-k)}{B(\alpha, \beta)}$$

Интересно, что знаменатель $P(X)$ в данном случае является самостоятельным *дискретным* распределением, называемым **бета-биномиальным распределением** (beta-binomial distribution)

$$BetaBinom(k | n, \alpha, \beta) = \binom{n}{k} \frac{B(\alpha+k, n+\beta-k)}{B(\alpha, \beta)},$$

показывающим вероятность k успехов в n испытаниях, при условии, что вероятность успеха p (она же θ) следует бета-распределению $Beta(\alpha, \beta)$.

Например, рассмотрим серию из $n = 6$ испытаний, в которых параметр p следует $Beta(2, 2)$. Вначале построим график бета-распределения (листинг 68 и рисунок 69).

```
plt.plot(theta, beta.pdf(theta, 2, 2), label = r'')
plt.title('Beta (2, 2)')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')
plt.show()
```

Листинг 68: $Beta(2, 2)$

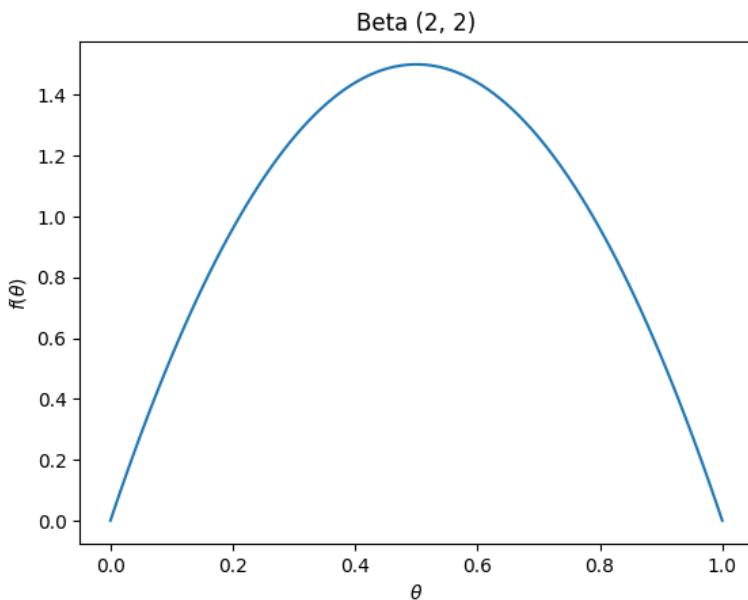


Рис. 69: $Beta(2, 2)$

Теперь выведем соответствующее бета-биномиальное распределение (листинг 69 и рисунок 70).

```
from scipy.stats import betabinom

n, a, b = 6, 2, 2
k = np.arange(n+1)

plt.plot(k, betabinom.pmf(k, n, a, b), 'bo', ms=8)
plt.vlines(k, 0, betabinom.pmf(k, n, a, b), colors = 'b', lw = 5, alpha
           =0.5)
plt.title('Beta-binomial distribution')
plt.xlabel('k')
plt.ylabel('f(k | 6, 2, 2)')
plt.show()
```

Листинг 69: $BetaBin(6, 2, 2)$

Согласно графику наиболее вероятным при таких параметрах будет $k = 3$ успехам. Убедимся в этом, воспользовавшись формулой матожидания бета-биномиального распределения:

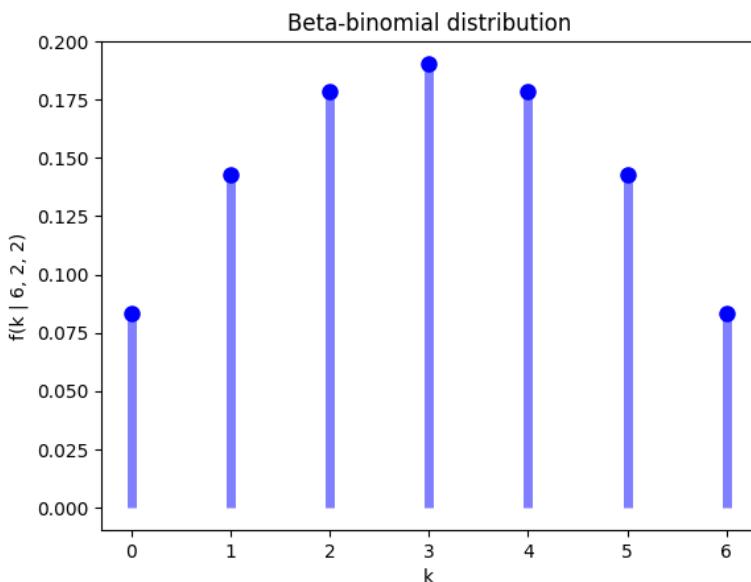


Рис. 70: $BetaBin(6, 2, 2)$

ленияя.

$$\mathbb{E} = \frac{n\alpha}{\alpha + \beta} = \frac{6 \cdot 2}{2 + 2} = 3$$

Если же речь идет о $Beta(1, 1)$, то есть непрерывном равномерном распределении, в котором все варианты p равновозможны, то бета-биномиальное распределение примет вид дискретного равномерного распределения (листинг 70 и рисунок 71).

```
n, a, b = 6, 1, 1
k = np.arange(n+1)

plt.plot(k, betabinom.pmf(k, n, a, b), 'bo', ms=8)
plt.vlines(k, 0, betabinom.pmf(k, n, a, b), colors = 'b', lw = 5, alpha
=0.5)
plt.title('Beta-binomial distribution')
plt.xlabel('k')
plt.ylabel('f(k | 6, 1, 1)')
plt.show()
```

Листинг 70: $BetaBin(6, 1, 1)$

Вернемся к формуле Байеса.

Апостериорная вероятность. Найдем апостериорную вероятность.

$$P(\theta | X) = \frac{P(\theta, X)}{P(X)} = \frac{\binom{n}{k} \frac{\theta^{\alpha+k-1} (1-\theta)^{n+\beta-k-1}}{B(\alpha, \beta)}}{\binom{n}{k} \frac{B(\alpha+k, n+\beta-k)}{B(\alpha, \beta)}},$$

где $B(\alpha, \beta)$ — нормализующая константа априорного распределения, а $B(\alpha+k, n+\beta-k)$ — нормализующая константа соответственно апостериорного распределения. Упростим

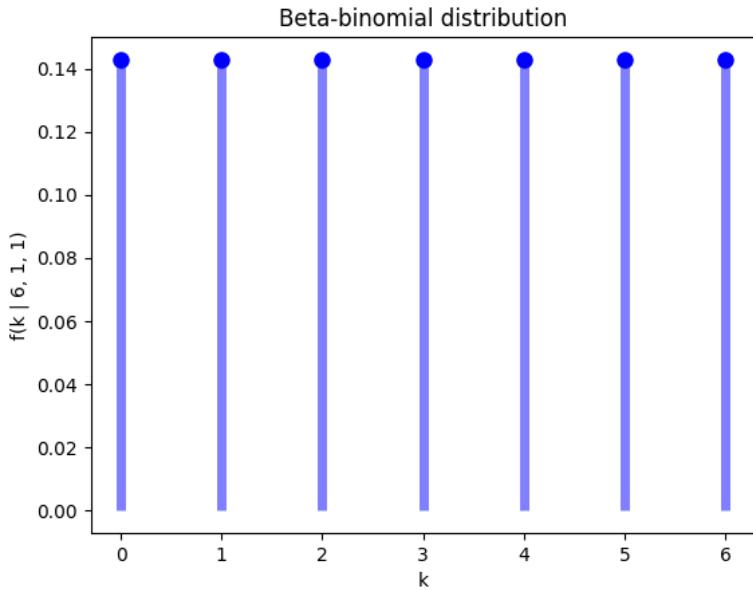


Рис. 71: $BetaBin(6, 1, 1)$

выражение.

$$\begin{aligned} P(\theta | X) &= \frac{\theta^{\alpha+k-1}(1-\theta)^{n+\beta-k-1}}{B(\alpha+k, n+\beta-k)} \\ &= \frac{\theta^{\alpha'-1}(1-\theta)^{\beta'-1}}{B(\alpha', \beta')} = Beta(\alpha', \beta') \end{aligned}$$

Распространенность заболевания. Вернемся к примеру с медицинскими тестами и распространенностью заболевания. Возможно уже в самом начале вам хотелось посчитать не вероятность заболевания при условии положительного теста, а его распространенность (априорное знание) с учетом того, сколько человек из выборки заболели (данные).

Раньше нам не удавалось этого сделать, так как априорная случайная величина была дискретной. Теперь, зная о сопряженности бета- и биномиального распределений, эта задача вполне выполнима.

Предположим, что больным оказался один человек из десяти, а о распространенности заболевания у нас нет никакой предварительной информации (в этом случае еще говорят об объективной или неинформативной априорной оценке). Тогда,

$$P(X | \theta) = \binom{10}{1} \theta^1 (1-\theta)^9, \quad P(\theta) = Beta(1, 1)$$

$$P(\theta | X) = Beta(\alpha + k, \beta + n - k) = Beta(1 + 1, 1 + 10 - 1) = Beta(2, 10)$$

Приведем график (листинг 71 и рисунок 72).

```

a, b = 1, 1
k, n = 1, 10

prior = beta.pdf(theta, a, b)
posterior = beta.pdf(theta, a+k, b+n-k)

plt.plot(theta, prior, label = 'prior, Beta(1,1)')
plt.plot(theta, posterior, label = 'posterior, Beta(2,10)')

plt.vlines(x = k/n, ymin = 0, ymax = 4.23, colors = 'r', label = r'$\theta_{MAP}$')

plt.title('Beta prior and posterior distributions')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')

plt.legend()
plt.show()

```

Листинг 71: Распространенность заболевания

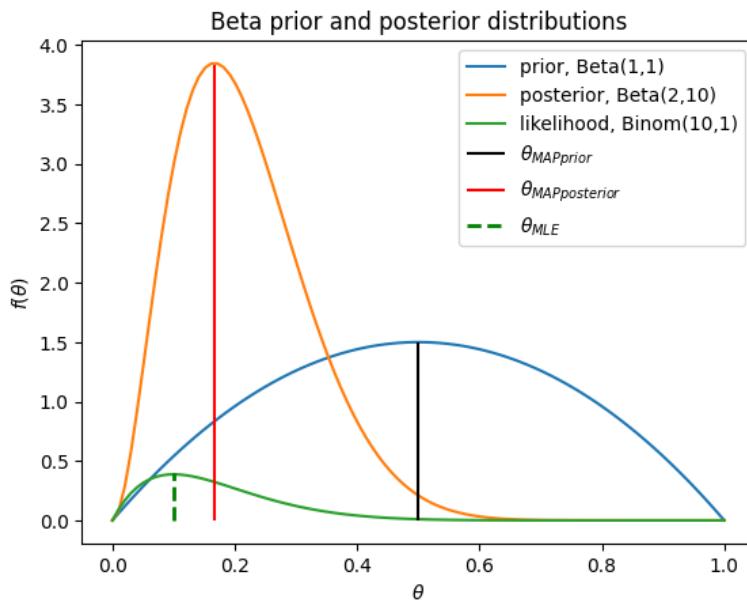


Рис. 72: Распространенность заболевания

Обратите внимание, что апостериорный максимум θ_{MAP} мы нашли через формулу максимального правдоподобия биномиального распределения $\theta_{MLE} = \frac{k}{n}$. Посмотрим, как быть, если априорное распределение неравномерно.

5.4 Мода априорного и апостериорного распределений

Априорным и апостериорным максимумом будут моды соответствующих распределений. Тогда априорный максимум бета-распределения можно найти по формуле

$$Beta(\alpha, \beta)_{mode} = \frac{(\alpha - 1)}{(\alpha - 1) + (\beta - 1)}$$

Апостриорный максимум в этом случае находится по формуле

$$\begin{aligned}\theta_{MAP} &= Beta(k + \alpha, n - k + \beta)_{mode} \\ &= \frac{k + \alpha - 1}{(k + \alpha - 1) + (n - k + \beta - 1)} \\ &= \frac{k + (\alpha - 1)}{n + (\alpha - 1) + (\beta - 1)}\end{aligned}$$

Разберем формулу апостериорного максимума более внимательно. Преобразуем выражение.

$$\theta_{MAP} = \frac{k}{n + (\alpha - 1) + (\beta - 1)} + \frac{\alpha - 1}{n + (\alpha - 1) + (\beta - 1)}$$

Умножим первое слагаемое на $\frac{n}{n}$, а второе на $\frac{(\alpha-1)+(\beta-1)}{(\alpha-1)+(\beta-1)}$. Тогда,

$$\theta_{MAP} = \underbrace{\frac{k}{n}}_{\theta_{MLE}} \cdot \underbrace{\frac{n}{n + (\alpha - 1) + (\beta - 1)}}_{w_1(\alpha, \beta, n)} + \underbrace{\frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}}_{Beta(\alpha, \beta)_{mode}} \cdot \underbrace{\frac{(\alpha - 1) + (\beta - 1)}{n + (\alpha - 1) + (\beta - 1)}}_{w_2(\alpha, \beta, n)}$$

Как мы видим, апостериорный максимум θ_{MAP} представляет собой средневзвешенное значение максимального правдоподобия θ_{MLE} и априорного максимума $Beta(\alpha, \beta)_{mode}$. Веса при этом зависят от α, β, n ,

- чем больше α, β , то есть чем больше наша уверенность в априорном знании, тем больше числитель и вся дробь w_2 и тем меньше дробь w_1 , а значит больше значимость априорного распределения; и наоборот
- чем больше n , то есть чем больше собрано данных, тем больше w_1 и тем меньше w_2 , а значит больше значимость правдоподобия.

Аналогичное утверждение справедливо для матожидания апостериорного распределения.

$$\mathbb{E}(\theta | X) = \frac{k + \alpha}{n + \alpha + \beta} = \frac{k}{n + \alpha + \beta} + \frac{\alpha}{n + \alpha + \beta}$$

Умножим первое слагаемое на $\frac{n}{n}$, а второе на $\frac{\alpha+\beta}{\alpha+\beta}$. Тогда,

$$\mathbb{E}(\theta | X) = \underbrace{\frac{k}{n}}_{\theta_{MLE}} \cdot \underbrace{\frac{n}{n + \alpha + \beta}}_{w_1(\alpha, \beta, n)} + \underbrace{\frac{\alpha}{\alpha + \beta}}_{\mathbb{E}(\theta)} \cdot \underbrace{\frac{\alpha + \beta}{n + \alpha + \beta}}_{w_2(\alpha, \beta, n)}$$

Закрепим на примере (листинги 72 и 73, а также рисунок 73). Вновь вычислим распространенность заболевания, но уже с априорным распределением $Beta(2, 2)$.

```

from math import comb

a, b = 2, 2
k, n = 1, 10

theta_max_prior = (a-1)/((a-1)+(b-1))
max_likelihood = k/n
theta_max_posterior = (k+(a-1)) / (n+(a-1)+(b-1))

theta_max_prior, max_likelihood, theta_max_posterior

Output: (0.5, 0.1, 0.1666666666666666)

```

Листинг 72: Моды распределений

```

prior = beta.pdf(theta, a, b)
posterior = beta.pdf(theta, a+k, b+n-k)
likelihood = comb(10, 1) * theta**k * (1-theta)**(n-k)

plt.plot(theta, prior, label = 'prior, Beta(2,2)')
plt.plot(theta, posterior, label = 'posterior, Beta(2,10)')
plt.plot(theta, likelihood, label = 'likelihood, Binom(10,1)')

plt.vlines(x = theta_max_prior, ymin = 0, ymax = 1.5, colors = 'k',
           label = r'$\theta_{MAP \ prior}$')
plt.vlines(x = theta_max_posterior, ymin = 0, ymax = 3.83, colors = 'r',
           label = r'$\theta_{MAP \ posterior}$')
plt.vlines(x = max_likelihood, ymin = 0, ymax = 0.4, colors = 'g',
           linestyles = '--', lw = 2, label = r'$\theta_{MLE}$')

plt.title('Beta prior and posterior distributions')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')

plt.legend()
plt.show()

```

Листинг 73: Моды распределений распределений на графике

И действительно, как мы видим, мода апостериорного распределения оказалась между модой априорного распределения и максимальным правдоподобием. Так как данных собрано относительно много, правдоподобие «перевесило» априорные убеждения.

5.5 Прогнозное априорное распределение

Как мы сказали ранее, априорное распределение является нашим представлением о распределении параметра θ до того, как мы собрали данные.

Возникает вопрос, можем ли мы предположить, *каким будет распределение данных X до того, как мы их получили*. Такое предположение называется **прогнозным априорным распределением** (prior predictive distribution).

Можно также сказать, что это ожидаемое (совместное) распределение данных с учетом

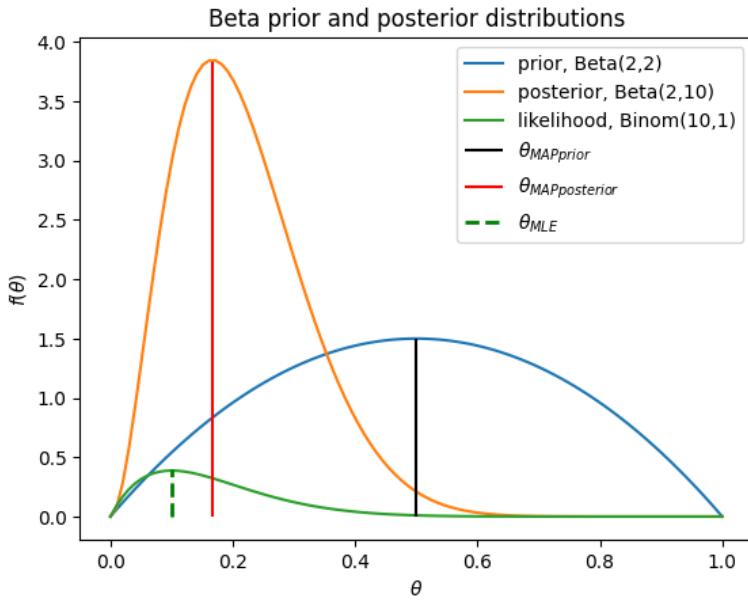


Рис. 73: Моды распределений

всех возможных значений $\theta \in \Theta$, оно же формула полной вероятности или знаменатель формулы Байеса $P(X)$.

$$P(X) = \int_{\theta \in \Theta} P(X, \theta) d\theta = \int_{\theta \in \Theta} P(X | \theta) \cdot P(\theta) d\theta$$

Как мы помним, у сопряженных априорного бета-распределения и биномиального правдоподобия знаменатель представляет собой бета-биномиальное распределение, которое как раз и является распределением X с учетом всех возможных значений $\theta \in [0, 1]$ и априорного знания о θ через параметры α, β .

$$P(X) = BetaBinom(k | n, \alpha, \beta) = \binom{n}{k} \frac{B(\alpha + k, n + \beta - k)}{B(\alpha, \beta)}$$

Построим бета-биномиальное распределение для примера распространенности заболевания с априорным распределением $Beta(2, 2)$ и ожидаемым количеством обследованных людей $n = 10$ (листинг 74 и рисунок 74).

```
n, a, b = 10, 2, 2
k = np.arange(n+1)

plt.plot(k, betabinom.pmf(k, n, a, b), 'bo', ms=8)
plt.vlines(k, 0, betabinom.pmf(k, n, a, b), colors = 'b', lw = 5, alpha
=0.5)
plt.title('Prior predictive distribution')
plt.xlabel('k')
plt.ylabel('f(k | 10, 2, 2)')
plt.show()
```

Листинг 74: Прогнозное априорное распределение

Поясним. Изначально мы считали, что распространенность составляет около 50-ти процентов, что отражено в распределении $Beta(2, 2)$. Как следствие, мы ожидаем, что с наибольшей вероятностью мы обнаружим заболевшими пять человек из десяти (листинг 75).

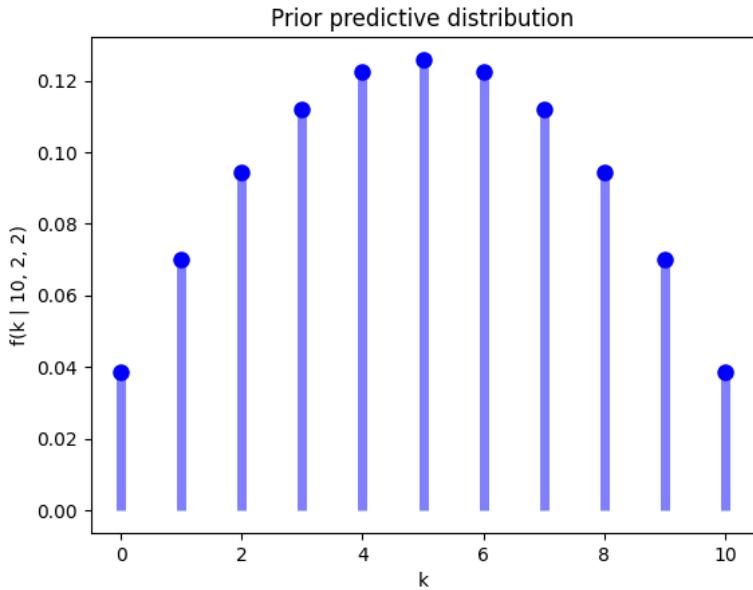


Рис. 74: Прогнозное априорное распределение

```
betabinom.mean(n, a, b)
```

```
Output: 5.0
```

Листинг 75: Среднее значение прогнозного априорного распределения

Конечно, так как в данном случае мы выбрали не вполне точное априорное распределение, полученное прогнозное априорное распределение далеко от фактически собранных данных. Напомню, что в выборке заболевшим был только один человек из десяти, в прогнозном же распределении вероятность того, что заболело не более одного человека равна лишь примерно 11-ти процентам (листинг 76).

```
betabinom.cdf(1, n, a, b)
```

```
Output: 0.10839160839160839
```

Листинг 76: Вероятность заболевания не более одного человека

5.6 Прогнозное апостериорное распределение

Аналогичным образом мы можем сделать предположение относительно того, каким будет распределение новых еще не знакомых нам данных X' после того, как мы высказали наше априорное предположение $P(\theta)$ и учли правдоподобие $P(X | \theta)$ уже собранных данных X .

Обозначим такое распределение через $P(X' | X)$ и назовем **прогнозным апостериорным распределением** (posterior predictive distribution).

$$P(X' | X) = \int_{\theta \in \Theta} P(X', \theta | X) d\theta = \int_{\theta \in \Theta} P(X' | \theta, X) \cdot P(\theta | X) d\theta$$

Здесь $P(\theta | X)$ представляет собой апостериорное (новое априорное) распределение. В случае же $P(X' | \theta, X)$ вспомним, что правдоподобие данных X' зависит только от θ , но

никак не от X . Тогда,

$$P(X' | \theta, X) = P(X' | \theta)$$

$$P(X' | X) = \int_{\theta \in \Theta} P(X' | \theta) \cdot P(\theta | X) d\theta$$

Для сопряженных бета- и биномиального распределений, если ввести новые параметры $\alpha' = \alpha + k$ и $\beta' = n + \beta - k$, а новые испытания и успехи в этих испытаниях обозначить n^* и k^* соответственно, то

$$P(X' | X) = \text{BetaBinom}(k^* | n^*, \alpha', \beta')$$

$$= \binom{n^*}{k^*} \frac{B(\alpha' + k^*, n^* + \beta' - k^*)}{B(\alpha', \beta')}$$

$$= \binom{n^*}{k^*} \frac{B(\alpha + k + k^*, n^* + n + \beta - k - k^*)}{B(\alpha + k, n + \beta - k)}$$

Посмотрим, каким будет прогнозное апостериорное распределение данных в нашем примере (листинг 77 и рисунок 75).

```
a, b = 2, 2
n, k = 10, 1
a_prime, b_prime = a + k, n + b - k
n_ast = 10
k_ast = np.arange(n_ast+1)

plt.plot(k_ast, betabinom.pmf(k_ast, n_ast, a_prime, b_prime), 'bo', ms=8)
plt.vlines(k_ast, 0, betabinom.pmf(k_ast, n_ast, a_prime, b_prime),
           colors='b', lw=5, alpha=0.5)
plt.title('Posterior predictive distribution')
plt.xlabel(r'$k^{\ast}$')
plt.ylabel(r'$f(k^{\ast}) \mid 10, 10 + 2, 10 + 2 - 1$')
plt.show()
```

Листинг 77: Прогнозное апостериорное распределение

Ожидаемо, после того как мы учли уже собранные данные X , прогнозное распределение новых данных X' сильно изменилось.

6 Сопряженные нормальные распределения

Рассмотрим нормальное априорное/апостериорное распределение и нормально распределенное правдоподобие. Поставим задачу построить распределение роста мужчин в России.

6.1 Правдоподобие

Как мы знаем, рост следует нормальному распределению, а значит мы можем предположить, что получив данные n респондентов, их рост будет распределен согласно

$$P(X | \theta) \sim \mathcal{N}(\theta, \sigma_x^2)$$

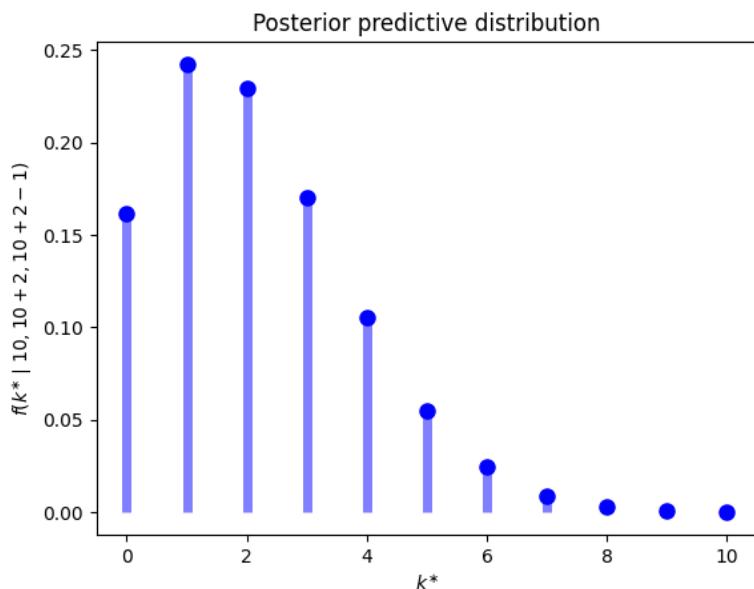


Рис. 75: Прогнозное апостериорное распределение

Для простоты предположим, что дисперсия σ_x^2 собранных данных не меняется. Например, мы знаем, что такие исследования проводятся каждый год, и дисперсия каждый год примерно одна и та же.

Кроме того, положим, что x_i является одним из x_1, x_2, \dots, x_n наблюдений в выборке X . Тогда $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ — выборочное среднее за текущий год (год, когда собирались данные).

6.2 Априорное распределение

Исходя из исследований за предыдущие годы мы знаем, что средний рост распределен согласно

$$P(\theta) \sim \mathcal{N}(\theta_0, \sigma_\theta^2),$$

где θ_0 — среднее значение роста за предыдущие годы, а σ_θ^2 — дисперсия среднего роста.

Еще раз обратим внимание, что

- σ_x^2 — дисперсия роста в выборке X ; в то время как
- σ_θ^2 — дисперсия среднего.

6.3 Апостериорное распределение

Покажем, что апостериорное распределение также нормально распределено

$$P(\theta | X) \sim \mathcal{N}(\theta', \sigma'^2)$$

и найдем аналитическое решение для апостериорных параметров θ' и σ'^2 . Как мы знаем,

$$P(\theta | X) \propto P(X | \theta) \cdot P(\theta)$$

Займемся правдоподобием. В первую очередь вспомним, что если у нас x_1, x_2, \dots, x_n независимых наблюдений (а мы предполагаем, что рост одного человека никак не зависит от роста других людей), то их совместная вероятность является произведением вероятностей отдельных наблюдений.

$$P(X = x_1, x_2, \dots, x_n \mid \theta) = \prod_{i=1}^n P(X = x_i \mid \theta)$$

Тогда, согласно нашему предположению о нормальном распределении собранных данных, оно равно

$$P(X \mid \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma_x^2}\right)$$

Примечание. Нотация $\exp(a)$ идентична записи e^a .

Так как первый множитель $\frac{1}{\sqrt{2\pi\sigma_x^2}}$ — константа, то

$$P(X \mid \theta) \propto \prod_{i=1}^n \exp\left(-\frac{(x_i - \theta)^2}{2\sigma_x^2}\right)$$

Кроме того, поскольку $e^a \cdot e^b = e^{a+b}$, то

$$P(X \mid \theta) \propto \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma_x^2}\right)$$

Перейдем к априорному распределению.

$$\begin{aligned} P(\theta) &= \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_\theta^2}\right) \\ &\propto \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_\theta^2}\right) \end{aligned}$$

Таким образом, апостериорное распределение будет иметь вид

$$P(\theta \mid X) \propto \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma_x^2}\right) \cdot \exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_\theta^2}\right)$$

Обратимся к первому множителю. По формуле квадрата разности.

$$\exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma_x^2}\right) = \exp\left(-\frac{\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2)}{2\sigma_x^2}\right)$$

Поскольку, $\sum(a + b + c) = \sum a + \sum b + \sum c$, то

$$\begin{aligned} &= \exp\left(-\frac{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n 2x_i\theta - \sum_{i=1}^n \theta^2}{2\sigma_x^2}\right) \\ &= \exp\left(-\frac{-\sum_{i=1}^n x_i^2 + 2\theta \sum_{i=1}^n x_i - \sum_{i=1}^n \theta^2}{2\sigma_x^2}\right) \end{aligned}$$

Заметим, что

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \sum_{i=1}^n x_i = n\bar{x}$$

$$\sum_{i=1}^n \theta^2 = n\theta^2$$

Последнее справедливо, так как мы складываем θ^2 n раз. Тогда,

$$\exp\left(\frac{-\sum_{i=1}^n x_i^2 + 2\theta n\bar{x} - n\theta^2}{2\sigma_x^2}\right)$$

Рассмотрим второй множитель.

$$\exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_\theta^2}\right) = \exp\left(-\frac{\theta^2 - 2\theta\theta_0 + \theta_0^2}{2\sigma_\theta^2}\right)$$

Подставим результаты в формулу апостериорной вероятности.

$$P(\theta | X) \propto \exp\left(\frac{-\sum_{i=1}^n x_i^2 + 2\theta n\bar{x} - n\theta^2}{2\sigma_x^2}\right) \cdot \exp\left(-\frac{\theta^2 - 2\theta\theta_0 + \theta_0^2}{2\sigma_\theta^2}\right)$$

По свойству $e^a \cdot e^b = e^{a+b}$,

$$P(\theta | X) \propto \exp\left[\frac{-\sum_{i=1}^n x_i^2 + 2\theta n\bar{x} - n\theta^2}{2\sigma_x^2} - \frac{\theta^2 - 2\theta\theta_0 + \theta_0^2}{2\sigma_\theta^2}\right]$$

Так как речь идет о пропорциональности, а не о равенстве, мы можем опустить константы, то есть те слагаемые, которые не зависят от θ .

$$\begin{aligned} P(\theta | X) &\propto \exp\left[\frac{-\sum_{i=1}^n x_i^2 + 2\theta n\bar{x} - n\theta^2}{2\sigma_x^2} - \frac{\theta^2 - 2\theta\theta_0 + \theta_0^2}{2\sigma_\theta^2}\right] \\ &\propto \exp\left[\frac{2\theta n\bar{x} - n\theta^2}{2\sigma_x^2} - \frac{\theta^2 - 2\theta\theta_0}{2\sigma_\theta^2}\right] \end{aligned}$$

Перепишем выражение.

$$P(\theta | X) \propto \exp\left[\frac{2\theta n\bar{x}}{2\sigma_x^2} - \frac{n\theta^2}{2\sigma_x^2} - \frac{\theta^2}{2\sigma_\theta^2} - \frac{2\theta\theta_0}{2\sigma_\theta^2}\right]$$

Сгруппируем слагаемые, содержащие θ^2 и θ .

$$P(\theta | X) \propto \left[-\frac{\theta^2}{2} \left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_x^2}\right) + \theta \left(\frac{\theta_0}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right)\right]$$

Положим

$$\begin{aligned} \sigma_\theta^{2'} &= \left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_x^2}\right)^{-1} \\ \theta' &= \sigma_\theta^{2'} \left(\frac{\theta_0}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right) \end{aligned}$$

Тогда,

$$P(\theta | X) \propto \exp\left[-\frac{\theta^2}{2\sigma_\theta^{2'}} + \frac{\theta\theta'}{\sigma_\theta^{2'}}\right]$$

Умножим второе слагаемое на $\frac{2}{2}$. Кроме того, так как речь идет о пропорциональности, мы можем добавить константу, т.е. выражение, не зависящее от θ , $\frac{\theta'^2}{2\sigma_\theta^{2'}}$.

$$P(\theta | X) \propto \exp\left[-\frac{\theta^2}{2\sigma_\theta^{2'}} + \frac{2\theta\theta'}{2\sigma_\theta^{2'}} - \frac{\theta'^2}{2\sigma_\theta^{2'}}\right]$$

Перед нами квадрат разности.

$$P(\theta | X) \propto \exp \left[-\frac{(\theta - \theta')^2}{2\sigma_{\theta}^{2'}} \right]$$

Для того чтобы превратить это выражение в полноценное распределение, добавим константу.

$$P(\theta | X) = \frac{1}{\sqrt{2\pi\sigma_{\theta}^{2'}}} \exp \left[-\frac{(\theta - \theta')^2}{2\sigma_{\theta}^{2'}} \right] \sim \mathcal{N}(\theta', \sigma_{\theta}^{2'})$$

6.4 Апостериорные параметры

Дисперсия. Рассмотрим формулу апостериорной дисперсии $\sigma_{\theta}^{2'}$.

$$\sigma_{\theta}^{2'} = \left(\frac{1}{\sigma_{\theta}^2} + \frac{n}{\sigma_x^2} \right)^{-1}$$

Введем понятие precision τ .

$$\tau = \frac{1}{\sigma^2}$$

В то время как дисперсия показывает, насколько удалены данные от среднего значения, precision показывает насколько они близки к среднему. Тогда,

$$\tau' = \frac{1}{\sigma_{\theta}^{2'}} = \underbrace{\frac{1}{\sigma_{\theta}^2}}_{\tau_{prior}} + \underbrace{\frac{n}{\sigma_x^2}}_{\tau_X}$$

Другими словами, апостериорная precision состоит из априорной precision τ_{prior} и precision собранных данных τ_X . Получается, что

- с уменьшением априорной дисперсии σ_{θ}^2 увеличивается априорная precision τ_{prior} и больший вес отдается априорному знанию;
- с увеличением количества данных n увеличивается precision данных τ_X ;
- аналогично, если уменьшается дисперсия данных, это также приводит к росту τ_X .

Матожидание. Подставим значение $\sigma_{\theta}^{2'}$ в формулу θ' и преобразуем выражение.

$$\begin{aligned} \theta' &= \sigma_{\theta}^{2'} \left(\frac{\theta_0}{\sigma_{\theta}^2} + \frac{n\bar{x}}{\sigma_x^2} \right) = \frac{\left(\frac{\theta_0}{\sigma_{\theta}^2} + \frac{n\bar{x}}{\sigma_x^2} \right)}{\left(\frac{1}{\sigma_{\theta}^2} + \frac{n}{\sigma_x^2} \right)} = \frac{\left(\frac{\theta_0\sigma_x^2 + n\bar{x}\sigma_{\theta}^2}{\sigma_{\theta}^2\sigma_x^2} \right)}{\left(\frac{\sigma_x^2 + n\sigma_{\theta}^2}{\sigma_{\theta}^2\sigma_x^2} \right)} = \frac{\theta_0\sigma_x^2 + n\bar{x}\sigma_{\theta}^2}{\sigma_x^2 + n\sigma_{\theta}^2} \\ &= \frac{\theta_0\sigma_x^2}{\sigma_x^2 + n\sigma_{\theta}^2} + \frac{n\bar{x}\sigma_{\theta}^2}{\sigma_x^2 + n\sigma_{\theta}^2} = \theta_0 \cdot \underbrace{\frac{\sigma_x^2}{\sigma_x^2 + n\sigma_{\theta}^2}}_{w_1} + \bar{x} \cdot \underbrace{\frac{n\sigma_{\theta}^2}{\sigma_x^2 + n\sigma_{\theta}^2}}_{w_2} \end{aligned}$$

Матожидание апостериорного распределения учитывает взвешенное по w_1 априорное среднее θ_0 и взвешенное по w_2 выборочное среднее \bar{x} . В частности,

- при уменьшении априорной дисперсии σ_{θ}^2 (т.е. большей уверенности в априорном знании) w_1 будет расти;
- при увеличении n , w_1 уменьшится, а w_2 наоборот увеличится;
- при уменьшении σ_x^2 , w_2 будет расти.

6.5 Прогнозное априорное распределение

Как мы уже знаем, для прогнозного априорного распределения нам нужно найти $P(X)$.

$$P(X) = \int_{-\infty}^{\infty} P(X, \theta) d\theta = \int_{-\infty}^{\infty} P(X | \theta) \cdot P(\theta) d\theta$$

В случае нормального правдоподобия и нормального априорного распределения

$$X \sim \mathcal{N}(\theta, \sigma_x^2)$$

$$\theta \sim \mathcal{N}(\theta_0, \sigma_\theta^2)$$

Тогда,

$$P(X) = \int_{-\infty}^{\infty} \mathcal{N}(\theta, \sigma_x^2) \cdot \mathcal{N}(\theta_0, \sigma_\theta^2) d\theta$$

В результате (приведем результат без доказательства) мы получим новое нормальное распределение с параметрами

$$P(X) = \mathcal{N}(\theta_0, \sigma_x^2 + \sigma_\theta^2)$$

И действительно, пока мы не собрали данные X , наилучшая оценка матожидания равна аналогичной априорной оценке θ_0 , дисперсия же является суммой дисперсии выборки σ_x^2 и априорной дисперсии среднего σ_θ^2 .

6.6 Прогнозное апостериорное распределение

Вспомним, что

$$P(X' | X) = \int_{-\infty}^{\infty} P(X', \theta | X) d\theta = \int_{-\infty}^{\infty} P(X' | \theta, X) \cdot P(\theta | X) d\theta$$

Опять же правдоподобие данных X' зависит только от θ , а не от X . Как следствие,

$$P(X' | X) = \int_{-\infty}^{\infty} P(X' | \theta) \cdot P(\theta | X) d\theta$$

При этом, так как дисперсия данных фиксирована

$$P(X' | \theta) \sim \mathcal{N}(\theta, \sigma_x^2)$$

$$P(\theta | X) = \mathcal{N}(\theta', \sigma_\theta^{2'})$$

По аналогии с прогнозным априорным распределением получим

$$P(X' | X) = \int_{-\infty}^{\infty} \mathcal{N}(\theta, \sigma_x^2) \cdot \mathcal{N}(\theta', \sigma_\theta^{2'}) d\theta = \mathcal{N}(\theta', \sigma_x^2 + \sigma_\theta^{2'})$$

Другими словами, пока мы не собрали новые данные X' , наилучшая оценка матожидания равна апостериорной оценке θ' , дисперсия же является суммой фиксированной дисперсии выборки σ_x^2 и апостериорной дисперсии $\sigma_\theta^{2'}$.

Пример. Проиллюстрируем сказанное выше на примере. Предположим, что мы собрали данные $n = 14$ респондентов, их средний рост составил $\bar{x} = 183$ см, дисперсия данных $\sigma_x^2 = 7$.

В предыдущих исследованиях средний рост составил $\theta_0 = 180$ см, дисперсия среднего $\sigma_\theta^2 = 1$.

Другими словами,

$$P(X | \theta) = \mathcal{N}(\theta, \sigma_x^2) = \mathcal{N}(\theta, 7)$$

$$P(\theta) = \mathcal{N}(\theta_0, \sigma_\theta^2) = \mathcal{N}(180, 1)$$

Найдем апостериорные параметры.

$$\sigma_{\theta'}^2 = \left(\frac{1}{\sigma_\theta^2} + \frac{n}{\sigma_x^2} \right)^{-1} = \left(\frac{1}{1} + \frac{14}{7} \right)^{-1} = \frac{1}{3}$$

$$\theta' = \sigma_{\theta'}^2 \left(\frac{\theta_0}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2} \right) = \frac{1}{3} \cdot \left(\frac{180}{1} + \frac{14 \cdot 183}{7} \right) = 182$$

Таким образом, апостериорное распределение имеет вид

$$P(\theta | X) = \mathcal{N}(\theta', \sigma_{\theta'}^2) = \mathcal{N}\left(182, \frac{1}{3}\right)$$

Посмотрим на графики распределений (листинг 78 и рисунок 76).

```
from scipy.stats import norm
prior = norm.pdf(theta, 180, np.sqrt(1))
posterior = norm.pdf(theta, 182, np.sqrt(1/3))

plt.plot(theta, prior, label = 'prior, Normal(180,1)')
plt.plot(theta, posterior, label = 'posterior, Normal(182,1/3)')

plt.title('Normal prior and posterior distributions')
plt.xlabel(r'$\theta$')
plt.ylabel(r'$f(\theta)$')

plt.legend()
plt.show()
```

Листинг 78: Сопряженные нормальные распределения

Также построим прогнозные априорное и апостериорное распределения (листинг 79 и рисунок 77).

$$P(X) = \mathcal{N}(\theta_0, \sigma_x^2 + \sigma_\theta^2) = \mathcal{N}(180, 7 + 1)$$

$$P(X' | X) = \mathcal{N}(\theta', \sigma_x^2 + \sigma_{\theta'}^2) = \mathcal{N}\left(182, 7 + \frac{1}{3}\right)$$

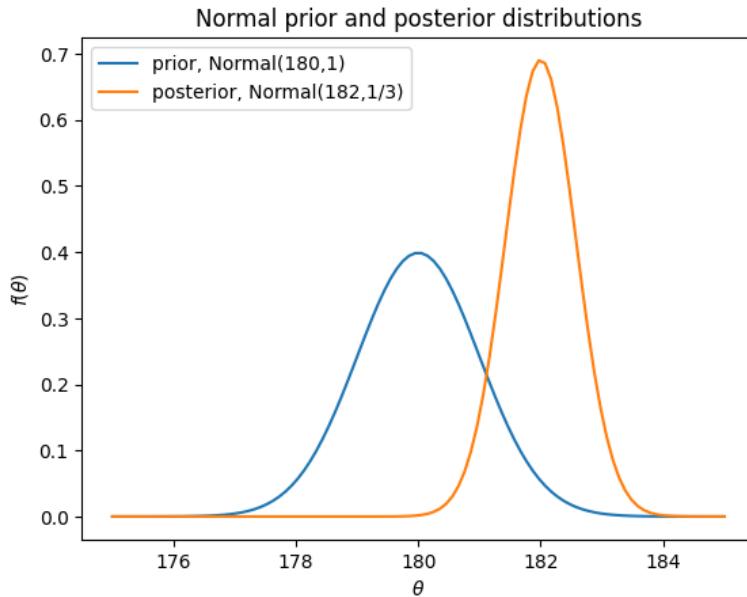


Рис. 76: Сопряженные нормальные распределения

```

theta = np.linspace(165, 195, 100)

prior_predictive = norm.pdf(theta, 180, np.sqrt(7+1))
posterior_predictive = norm.pdf(theta, 182, np.sqrt(7+1/3))

plt.plot(theta, prior_predictive, label = 'prior predictive')
plt.plot(theta, posterior_predictive, label = 'posterior predictive')

plt.title('Normal prior and posterior predictive distributions')
plt.xlabel(r'x')
plt.ylabel(r'$f(x)$')

plt.legend()
plt.show()

```

Листинг 79: Прогнозные априорное и апостериорное распределения

Продолжим изучать сопряженные распределения.

7 Процесс Пуассона

7.1 Распределение Пуассона

Предельный случай биномиального распределения. Ранее мы говорили, что согласно теореме Муавра-Лапласа при количестве испытаний n , стремящемся к бесконечности и вероятности успеха p близкой к 0,5, биномиальное распределение приближается к нормальному

$$\binom{n}{k} p^k q^{n-k} \sim \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right)$$

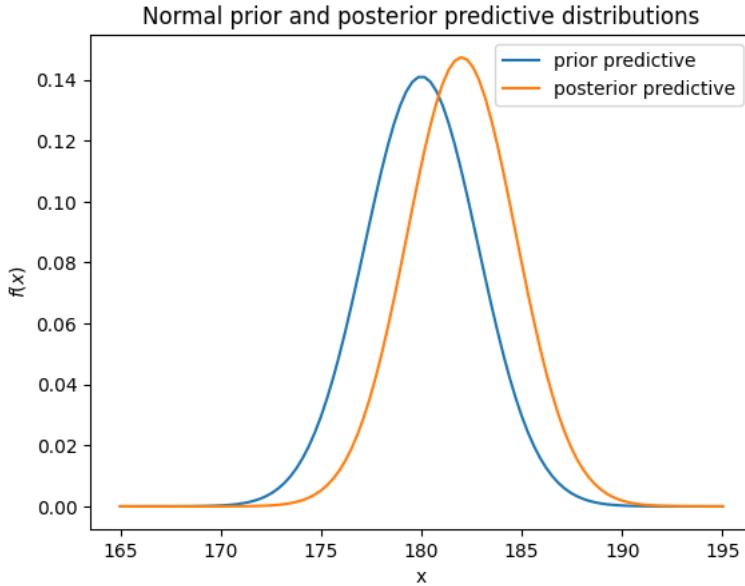


Рис. 77: Прогнозные априорное и апостериорное распределения

Другими словами, нормальное распределение является предельным случаем биномиального распределения. Еще одним предельным случаем биномиального распределения является распределение Пуассона при увеличении n и стремлении p к нулю. Приведем доказательство.

Предположим, что мы не знаем, сколько будет проведено испытаний или вероятность успеха в каждом из них. При этом пусть нам известно, что если проводить испытания в регулярные промежутки времени, то в среднем в них будет $\lambda = np$ успехов. Другими словами, предположим, что мы знаем матожидание биномиального распределения (ведь $\mathbb{E}[k] = np$), но не знаем конкретных параметров n и p .

Если $\lambda = np$, то $p = \frac{\lambda}{n}$ и $q = 1 - \frac{\lambda}{n}$. Перепишем функцию вероятности биномиального распределения, подставив эти выражения вместо p и q .

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n!}{(n-k)!} \cdot \frac{1}{k!} \cdot \frac{1}{n^k} \cdot \frac{\lambda^k}{1} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= \left(\frac{\lambda^k}{k!}\right) \cdot \left(\frac{n!}{(n-k)!} \cdot \frac{1}{n^k}\right) \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \end{aligned}$$

Найдем значение этого выражения при n , стремящемся к бесконечности. Тогда,

$$\lim_{n \rightarrow \infty} P(X = k) = \lim_{n \rightarrow \infty} \left(\frac{\lambda^k}{k!}\right) \cdot \left(\frac{n!}{(n-k)!} \cdot \frac{1}{n^k}\right) \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}$$

Первый множитель — это константа, которая не зависит от n .

$$\lim_{n \rightarrow \infty} \left(\frac{\lambda^k}{k!}\right) = \frac{\lambda^k}{k!}$$

Второй множитель можно переписать следующим образом.

$$\lim_{n \rightarrow \infty} \left(\frac{n!}{(n-k)!} \cdot \frac{1}{n^k}\right) = \lim_{n \rightarrow \infty} \left(\frac{n(n-1)(n-2)\dots(n-k)(n-k-1)\dots1}{(n-k)(n-k-1)\dots1}\right) \cdot \left(\frac{1}{n^k}\right)$$

Сократим общие множители.

$$= \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k}$$

Так как мы сократили $n - k$ множителей, то в числителе осталось k множителей от n до $n - k + 1$. В знаменателе, очевидно, также k множителей. Тогда

$$\lim_{n \rightarrow \infty} \left(\frac{n}{n} \right) \cdot \left(\frac{n-1}{n} \right) \cdot \left(\frac{n-2}{n} \right) \cdot \dots \cdot \left(\frac{n-k+1}{n} \right)$$

При стремлении n к бесконечности каждый из этих k множителей стремится к единице. Значит,

$$\lim_{n \rightarrow \infty} \left(\frac{n}{n} \right) \cdot \left(\frac{n-1}{n} \right) \cdot \left(\frac{n-2}{n} \right) \cdot \dots \cdot \left(\frac{n-k+1}{n} \right) = 1$$

Рассмотрим третий множитель.

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n} \right)^n$$

Как известно, число e можно определить как

$$e = \lim_{x \rightarrow \infty} \left(1 + \frac{1}{x} \right)^x$$

Если положить $x = -\frac{n}{\lambda}$, то $-\frac{\lambda}{n} = \frac{1}{x}$ и $n = x(-\lambda)$. Тогда,

$$\lim_{x \rightarrow \infty} \left(1 - \frac{\lambda}{n} \right)^n = \lim_{x \rightarrow \infty} \left(1 + \frac{1}{x} \right)^{x(-\lambda)} = e^{-\lambda}$$

Наконец последний множитель очевидно стремится к единице.

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n} \right)^{-k} = 1^{-k} = 1$$

Получаем распределение Пуассона

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

с носителем $k \in \{0, 1, 2, \dots\}$ и $\lambda \in (0, \infty)$.

Сумма вероятностей всех значений. Сумма вероятностей всех возможных (то есть от нуля до бесконечности) значений случайной величины k должна быть по определению равна единице, $\sum_{k=0}^{\infty} P(X = k) = 1$. Убедимся в этом.

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

Выпишем первые несколько слагаемых суммы.

$$\sum_{k=0}^{\infty} P(X = k) = e^{-\lambda} \left[\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right] = e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right]$$

Выражение в скобках — разложение экспоненциальной функции $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ в ряд Маклорена. А значит,

$$\sum_{k=0}^{\infty} P(X = k) = e^{-\lambda} \cdot e^{\lambda} = 1$$

Матожидание и дисперсия. Также найдем математическое ожидание и дисперсию.

Матожидание. По определению матожидания дискретной случайной величины

$$\mathbb{E}[X] = \sum xP(X=x)$$

Тогда для распределения Пуассона матожидание можно записать как

$$\mathbb{E}[k] = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!}$$

Первое слагаемое суммы при $k = 0$ обращается в ноль, $0 \frac{\lambda^0}{0!} = 0$, значит нас интересуют только слагаемые при $k \geq 1$.

$$\mathbb{E}[k] = e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!}$$

Начнем раскладывать сумму.

$$\mathbb{E}[k] = e^{-\lambda} \left[\lambda + \lambda^2 + \frac{\lambda^3}{2!} + \frac{\lambda^4}{3!} + \dots \right]$$

Вынесем λ за скобки.

$$\mathbb{E}[k] = \lambda e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \right]$$

Выражение в скобках — опять же разложение экспоненциальной функции $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ в ряд Маклорена. Следовательно,

$$\mathbb{E}[k] = \lambda \cdot e^{-\lambda} \cdot e^{\lambda} = \lambda$$

Дисперсия. Дисперсия распределения Пуассона также равна λ . Докажем это. Известно, что

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Вначале заметим, что второе слагаемое, т.е. квадрат матожидания k , как следует из доказательства выше, равен $\mathbb{E}[k] = \lambda \rightarrow \mathbb{E}[k]^2 = \lambda^2$. Найдем первое слагаемое, матожидание квадрата k , $\mathbb{E}[k^2]$.

Доказательство 1. Начнем с $\mathbb{E}[k(k-1)]$. Поскольку

$$\mathbb{E}[X(X-1)] = \sum x(x-1)P(X=x)$$

Получается, что

$$\mathbb{E}[k(k-1)] = \sum_{k=0}^{\infty} k(k-1) \cdot \frac{\lambda^k}{e^{-\lambda}}$$

При $k = 0$ и $k = 1$ члены ряда обращаются в ноль, следовательно

$$\begin{aligned} \mathbb{E}[k(k-1)] &= \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k \cdot e^{-\lambda}}{k!} = e^{-\lambda} \cdot \sum_{k=2}^{\infty} \cancel{k(k-1)} \frac{\lambda^k}{\cancel{k(k-1)(k-2)!}} \\ &= e^{-\lambda} \cdot \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} \end{aligned}$$

Поскольку $\lambda^k = \lambda^2(\lambda^{k-2})$, то

$$\mathbb{E}[k(k-1)] = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}$$

Если положить $z = k - 2$, тогда

$$\mathbb{E}[k(k-1)] = \lambda^2 e^{-\lambda} \sum_{z=0}^{\infty} \frac{\lambda^z}{z!}$$

Так как $\sum_{z=0}^{\infty} \frac{\lambda^z}{z!} = e^\lambda$, то

$$\begin{aligned}\mathbb{E}[k(k-1)] &= \lambda^2 \cdot e^{-\lambda} \cdot e^\lambda = \lambda^2 \\ \mathbb{E}[k(k-1)] &= \mathbb{E}[k^2] - \mathbb{E}[k] = \lambda^2\end{aligned}$$

Выразим и найдем $\mathbb{E}[k^2]$

$$\mathbb{E}[k^2] = \lambda^2 + \mathbb{E}[k] = \lambda^2 + \lambda$$

Отсюда

$$\mathbb{V}[k] = \mathbb{E}[k^2] - \mathbb{E}[k]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Матожидание квадрата случайной величины $\mathbb{E}[k^2]$ можно найти иначе.

Доказательство 2. Поскольку,

$$\mathbb{E}[X^2] = \sum x^2 P(X = x)$$

$$\mathbb{E}[k^2] = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}$$

Опять же первое слагаемое суммы при $k = 0$ обращается в ноль, значит нас интересуют только слагаемые при $k \geq 1$.

$$\begin{aligned}
\mathbb{E}[k^2] &= e^{-\lambda} \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k(k-1)!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{k \cdot \lambda \cdot \lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{k}{(k-1)!} \lambda^{k-1} \right) = \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{k-1+1}{(k-1)!} \lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{k-1}{(k-1)!} \lambda^{k-1} + \frac{1}{(k-1)!} \lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{k-1}{(k-1)!} \lambda^{k-1} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{k-1}{(k-1)(k-2)!} \lambda^{k-1} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{1}{(k-2)!} \lambda^{k-1} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{1}{(k-2)!} \lambda \cdot \lambda^{k-2} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \right) \\
&= \lambda e^{-\lambda} \left(\lambda \sum_{k=1}^{\infty} \frac{1}{(k-2)!} \lambda^{k-2} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \right)
\end{aligned}$$

Поскольку $\sum_{k=1}^{\infty} \frac{1}{(k-2)!} \lambda^{k-2}$ для $k < 2$ не определена, то

$$\mathbb{E}[k^2] = \lambda e^{-\lambda} \left(\lambda \sum_{k=2}^{\infty} \frac{1}{(k-2)!} \lambda^{k-2} + \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} \right)$$

Положив $i = k - 2$ и $j = k - 1$, получим

$$\mathbb{E}[k^2] = \lambda e^{-\lambda} \left(\lambda \sum_{i=0}^{\infty} \frac{1}{i!} \lambda^i + \sum_{j=0}^{\infty} \frac{1}{j!} \lambda^j \right)$$

Так как $\sum_{z=0}^{\infty} \frac{\lambda^z}{z!} = e^{\lambda}$, то

$$\mathbb{E}[k^2] = \lambda e^{-\lambda} (\lambda e^{\lambda} + e^{\lambda}) = \lambda e^{-\lambda} e^{\lambda} (\lambda + 1) = \lambda^2 + \lambda$$

Отсюда

$$\mathbb{V}[k] = \mathbb{E}[k^2] - \mathbb{E}[k]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Как следствие, поскольку $\mathbb{E}[k] = \mathbb{V}[k] = \lambda$, говорить о том, что данные следуют распределению Пуассона можно только в том случае, когда их среднее значение совпадает с дисперсией.

Мода распределения. При $k > 0$ отношение значений двух последовательных случайных величин равно

$$\frac{P(X = k)}{P(X = k - 1)} = \frac{e^{-\lambda} \frac{\lambda^k}{k!}}{e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!}} = \frac{e^{-\lambda} \frac{\lambda(\lambda^{k-1})}{k(k-1)!}}{e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!}} = \frac{\lambda}{k}$$

Рассмотрим случай, когда λ — **целое число** m . Пусть $\lambda = m = 4$, тогда

$$\begin{aligned} P(X = 0) &= \frac{4^0 \cdot e^{-4}}{0!} = e^{-4} & \frac{\lambda}{k} &= \frac{4}{1} \\ P(X = 1) &= \frac{4^1 \cdot e^{-4}}{1!} = \frac{4}{1} e^{-4} = 4e^{-4}, & \frac{\lambda}{k} &= \frac{4}{2} \\ P(X = 2) &= \frac{4^2 \cdot e^{-4}}{2!} = \frac{4^2}{2 \cdot 1} e^{-4} = 8e^{-4}, & \frac{\lambda}{k} &= \frac{4}{3} \\ P(X = 3) &= \frac{4^3 \cdot e^{-4}}{3!} = \frac{4^3}{3 \cdot 2 \cdot 1} e^{-4} = \frac{32}{3} e^{-4} \approx 10,7e^{-4}, & \frac{\lambda}{k} &= \frac{4}{4} \\ P(X = 4) &= \frac{4^4 \cdot e^{-4}}{4!} = \frac{4^4}{4 \cdot 3 \cdot 2 \cdot 1} e^{-4} = \frac{32}{3} e^{-4} \approx 10,7e^{-4}, & \frac{\lambda}{k} &= \frac{4}{5} \\ P(X = 5) &= \frac{4^5 \cdot e^{-4}}{5!} = \frac{4^5}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} e^{-4} = \frac{128}{15} e^{-4} \approx 8,53e^{-4}, & \frac{\lambda}{k} &= \frac{4}{6} \\ P(X = 6) &= \frac{4^6 \cdot e^{-4}}{6!} = \frac{4^6}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} e^{-4} = \frac{512}{90} e^{-4} \approx 5,69e^{-4}, & \frac{\lambda}{k} &= \frac{4}{6} \end{aligned}$$

Можно заметить, что соотношение $\frac{\lambda}{k}$ больше единицы вплоть до $P(X = 4)$, а затем — меньше. Более того, значения $P(X = m) = P(X = m-1)$ одинаковы и оба являются модой распределения. В данном случае модой будут значения $P(X = 4) = P(X = 4 - 1 = 3)$.

Более формально это можно выразить следующим образом. Поскольку,

$$P(X = k) = \frac{\lambda}{k} \cdot P(X = k - 1)$$

Значит

- если $k < \lambda$, $\frac{\lambda}{k} > 1$ и $P(X = k) > P(X = k - 1)$;
- если $k = \lambda$, $\frac{\lambda}{k} = 1$ и $P(X = k) = P(X = k - 1)$; и наконец
- если $k > \lambda$, $\frac{\lambda}{k} < 1$ и $P(X = k) < P(X = k - 1)$.

Если же λ — положительное **вещественное** число, то мода равна наибольшему целому числу, которое меньше λ . Такую зависимость можно выразить через функцию «пол», floor function, $\lfloor \lambda \rfloor$.

И действительно, пусть $\lambda = 4,1$, тогда

$$\begin{aligned}
 P(X=0) &= \frac{4,1^0 \cdot e^{-4,1}}{0!} = e^{-4,1} & \frac{\lambda}{k} &= \frac{4,1}{1} \\
 P(X=1) &= \frac{4,1^1 \cdot e^{-4,1}}{1!} = \frac{4,1}{1} e^{-4,1}, & \frac{\lambda}{k} &= \frac{4,1}{2} \\
 P(X=2) &= \frac{4,1^2 \cdot e^{-4,1}}{2!} = \frac{4,1^2}{2 \cdot 1} e^{-4,1}, & \frac{\lambda}{k} &= \frac{4,1}{3} \\
 P(X=3) &= \frac{4,1^3 \cdot e^{-4,1}}{3!} = \frac{4,1^3}{3 \cdot 2 \cdot 1} e^{-4,1}, & \frac{\lambda}{k} &= \frac{4,1}{4} \\
 P(X=4) &= \frac{4,1^4 \cdot e^{-4,1}}{4!} = \frac{4,1^4}{4 \cdot 3 \cdot 2 \cdot 1} e^{-4,1}, & \frac{\lambda}{k} &= \frac{4,1}{5} \\
 P(X=5) &= \frac{4,1^5 \cdot e^{-4,1}}{5!} = \frac{4,1^5}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} e^{-4,1}, & \frac{\lambda}{k} &= \frac{4,1}{6} \\
 P(X=6) &= \frac{4,1^6 \cdot e^{-4,1}}{6!} = \frac{4,1^6}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} e^{-4,1},
 \end{aligned}$$

Другими словами, $P(X = \lceil 4,1 \rceil) < P(X = \lfloor 4,1 \rfloor)$ или в общем случае

$$P(X = \lceil \lambda \rceil) < P(X = \lfloor \lambda \rfloor)$$

Правдоподобие. Таким образом, мы можем воспользоваться этим распределением для того, чтобы моделировать независимые события k , которые происходят с определенной интенсивностью λ за фиксированный промежуток времени или пространства.

$$\mathcal{L}(k | \lambda) \sim Pois(k | \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Примечание. Здесь и далее, следуя традиционной нотации для распределения Пуассона мы заменим параметр θ на λ .

Например, говоря о промежутках времени, мы можем построить распределение количества людей k , подходящих каждый час к кассе магазина.

Предположим, что в среднем каждый час к одной кассе подходят $\lambda = 6$ покупателей (листинг 80 и рисунок 78).

```

from scipy.stats import poisson

n, mu = 20, 6
k = np.arange(n+1)

plt.plot(k, poisson.pmf(k, mu), 'bo', ms=8)
plt.vlines(k, 0, poisson.pmf(k, mu), colors='b', lw=5, alpha=0.5)
plt.title('Poisson likelihood')
plt.xlabel('k')
plt.ylabel(r'$f(k | \lambda = 6)$')
plt.show()

```

Листинг 80: Распределение Пуассона как модель правдоподобия

Рассчитаем вероятность того, что в течение 30-ти минут придут 2 покупателя.

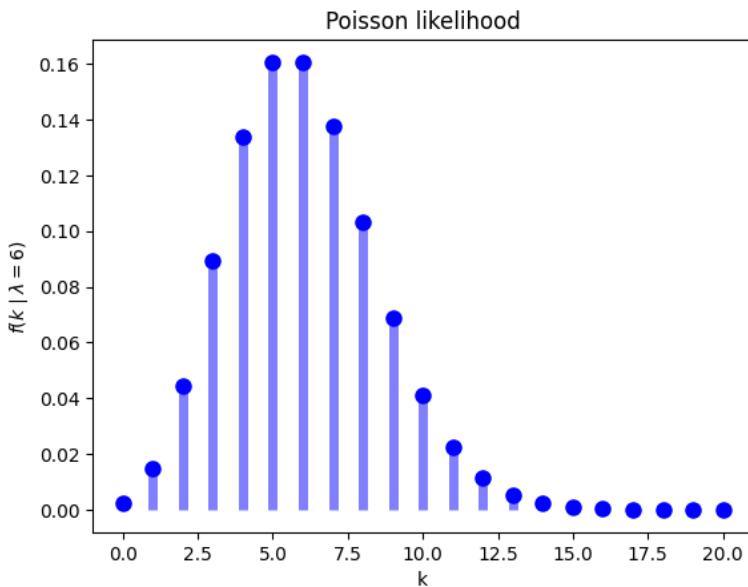


Рис. 78: Распределение Пуассона как модель правдоподобия

Здесь нам потребуется немного изменить формулу для того, чтобы учесть временной интервал t . Если за один час к кассе подходит λ покупателей, то за интервалом длительностью t подходит λt покупателей. Тогда,

$$P(k, t | \lambda) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

Поскольку 30 минут соответствуют $t = 0,5$ часа, получаем

$$P(2, 0,5 | 6) = \frac{(6 \cdot 0,5)^2 \cdot e^{(-6 \cdot 0,5)}}{2!} \approx 0,22$$

Сделаем расчет на Питоне (листинг 81).

```
lambda_, t = 6, 0.5
mu = lambda_ * t
poisson.pmf(2, mu)

Output: 0.22404180765538775
```

Листинг 81: Вероятность того, что в течение 30-ти минут придут 2 покупателя

Аналогично, в качестве фиксированной единицы пространства можно выбрать, например, несколько касс и с помощью распределения Пуассона моделировать вероятность подхода k покупателей к конкретной кассе.

Итак мы познакомились с распределением Пуассона и договорились использовать его в качестве функции правдоподобия. Начнем подбирать оптимальное априорное распределение. Однако прежде чем непосредственно перейти к рассмотрению априорного распределения, сделаем шаг назад и посмотрим в целом на то, какую ситуацию мы пытаемся моделировать и какую информацию получить.

7.2 Процесс (поток) Пуассона

В данном случае, мы исследуем процесс, в котором (1) есть независимые события и определенные *неодинаковые* временные интервалы T_i , через которые они происходят. Одновременно, нам известно, (2) сколько в среднем событий происходит за *фиксированный* интервал t , то есть λ . Такой процесс (рисунок 79) называется **процессом или потоком Пуассона** (Poisson process).

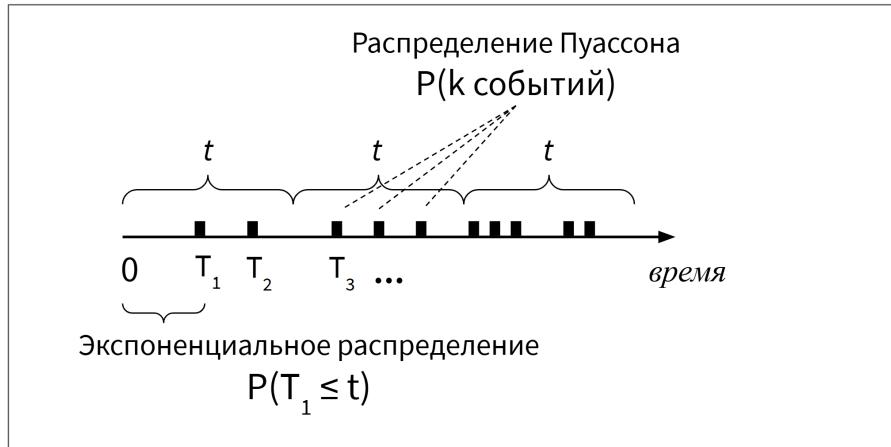


Рис. 79: Процесс (поток) Пуассона

Для того чтобы моделировать вероятность количества покупателей k , подходящих за равные промежутки времени t к кассе, можно использовать распределение Пуассона. Как это делается, мы уже рассмотрели.

7.3 Экспоненциальное распределение

Теперь попробуем оценить вероятность времени T_1 до первого события (в нашем примере, покупателя).

Вначале найдем вероятность того, что время T_1 до первого события больше фиксированного интервала t . Другими словами, вероятность того, что за время t не произойдет ни одного события. Такую вероятность можно найти с помощью распределения Пуассона при $k = 0$.

$$P(T_1 > t) = P(X = 0 \mid \lambda t) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

К такому же результату можно прийти, воспользовавшись формулой совместной вероятности независимых событий.

$$P(T_1 > t) = P(X = 0 \mid \lambda)^t = \left(\frac{\lambda^0 e^{-\lambda}}{0!} \right)^t = (e^{-\lambda})^t = e^{-\lambda t}$$

Одновременно, вероятность (хотя бы) одного события за время t равна

$$P(T_1 \leq t) = 1 - P(T_1 > t) = 1 - e^{-\lambda t}$$

Например, найдем вероятность того, что покупатель подойдет к кассе в течение трех минут. Здесь, $t = \frac{3}{60} = 0,05$ часа и $\lambda = 6$. Тогда,

$$P(T_1 \leq 0,05) = 1 - e^{(-6 \cdot 0,05)} \approx 0,26$$

Функция распределения и плотности вероятности. Интересно, что функция экспоненциального распределения (exponential distribution)

$$X \sim Exp(\lambda)$$

как раз имеет вид $cdf(t) = 1 - e^{-\lambda t}$, а значит описывает вероятность времени до первого события T_1 . Приведем код на Питоне (листинг 82).

```
from scipy.stats import expon
lambda_ = 6
b = 1/lambda_

expon.cdf(0.05, scale = b)

Output: 0.2591817793182822
```

Листинг 82: Вероятность времени до первого события как функция экспоненциального распределения

Примечание. Пояснения к используемым параметрам будут даны ниже.

Для того чтобы найти плотность вероятности, дифференцируем по t .

$$pdf(t) = \frac{d}{dt}(cdf) = \frac{d}{dt}(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$$

Построим график (листинг 83 и рисунок 80).

```
t = np.linspace(0, 1.5, 100)
plt.plot(t, expon.pdf(t, scale = b));
plt.title(r'Exponential distribution, $Exp(\lambda = 6)$')
plt.xlabel('t')
plt.ylabel('f(t)')
plt.show()
```

Листинг 83: Экспоненциальное распределение, $Exp(\lambda = 6)$

Посмотрим на вероятность того, что первый покупатель подойдет к кассе за один час (листинг 84 и рисунок 81).

```
t = np.linspace(0, 1.5, 100)
fill_t = np.linspace(0, 1.5, 100)
plt.plot(t, expon.pdf(t, scale = b))
plt.fill_between(fill_t, expon.pdf(fill_t, scale = b), alpha = 0.3)
plt.title(r'$Exp(\lambda = 6)$, $P(T_1 \leq 1)$')
plt.xlabel('t')
plt.ylabel('f(t)')
plt.show()
```

Листинг 84: Экспоненциальное распределение, $Exp(\lambda = 6)$, $P(T_1 \leq 1)$

Вычислим интеграл (листинг 85).

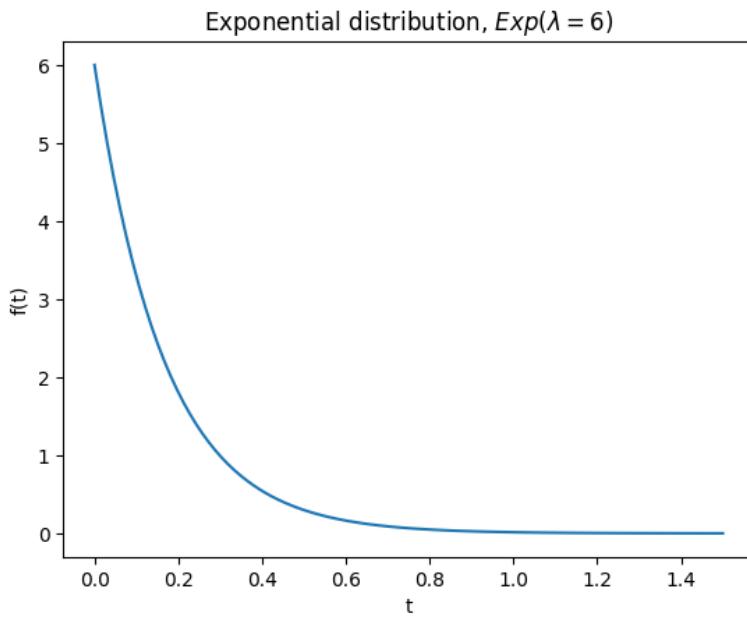


Рис. 80: Экспоненциальное распределение, $Exp(\lambda = 6)$

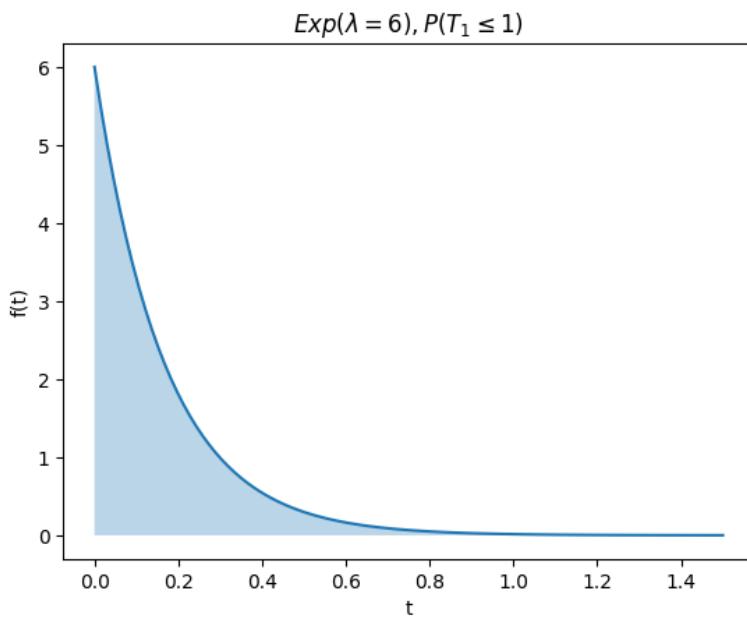


Рис. 81: Экспоненциальное распределение, $Exp(\lambda = 6)$, $P(T_1 \leq 1)$

```
from scipy.integrate import quad
quad(expon.pdf, 0, 1, args = (0, 1/6))[0]
Output: 0.997521247823336
```

Листинг 85: Вероятность $P(T_1 \leq 1)$

При интенсивности $\lambda = 6$ покупателей в час такая вероятность близка к единице. Теперь посмотрим (листинг 86) на вероятность T_1 через 10 минут, полчаса и час с помощью функции `expon.cdf()`.

```

time = [1/6, 1/2, 1]

for t_period in time:
    print(expon.cdf(t_period, scale = b))

Output:
0.6321205588285577
0.950212931632136
0.9975212478233336

```

Листинг 86: Вероятность T_1 через 10 минут, полчаса и час

Параметры распределения. Поясним параметры распределения.

Интенсивность и среднее время. Экспоненциальное (и как мы увидим некоторые другие распределения) могут определяться различными наборами параметров.

С одной стороны, как мы показали выше, мы можем задать экспоненциальное распределение через параметр λ , то есть **интенсивность** (rate) или среднее количество событий за фиксированный промежуток времени t . В примере выше, как мы не раз говорили, за один час к кассе в среднем подходит $\lambda = 6$ покупателей.

С другой стороны, и такая параметризация используется в библиотеке scipy, экспоненциальное распределение можно задать через $\beta = \frac{1}{\lambda}$, то есть **среднее время** (mean time, scale), за которое происходит одно событие. В нашем примере, один покупатель в среднем подходит к кассе каждую $\beta = \frac{1}{6}$ часа или каждые 10 минут.

Пороговое значение. Параметр θ (loc, **пороговое значение**, threshold) указывает на минимально возможное время до первого события.

$$cdf(t) = 1 - e^{-\lambda(t-\theta)}$$

$$pdf(t) = \lambda e^{-\lambda(t-\theta)}$$

Например, первый покупатель может подойти к кассе не ранее чем спустя три минуты (0,05 часа), поскольку, например, мы начинаем отсчет времени с момента, когда предыдущий покупатель подошел к этой кассе (и его обслуживание длится как раз три минуты).

Посмотрим на график (листинг 87 и рисунок 82).

```

t = np.linspace(0, 1.5, 100)
plt.plot(t, expon.pdf(t, loc = 0.05, scale = b))
plt.title(r'$Exp(\lambda = 6, \theta = 0.05)$')
plt.xlabel('t')
plt.ylabel('f(t)')
plt.show()

```

Листинг 87: Пороговое значение экспоненциального распределения, $Exp(\lambda = 6, \theta = 0.05)$

Логично, что вероятность того, что первый покупатель подойдет, скажем, через 10 минут, полчаса и час при $\theta > 0$ снизится (листинг 88).

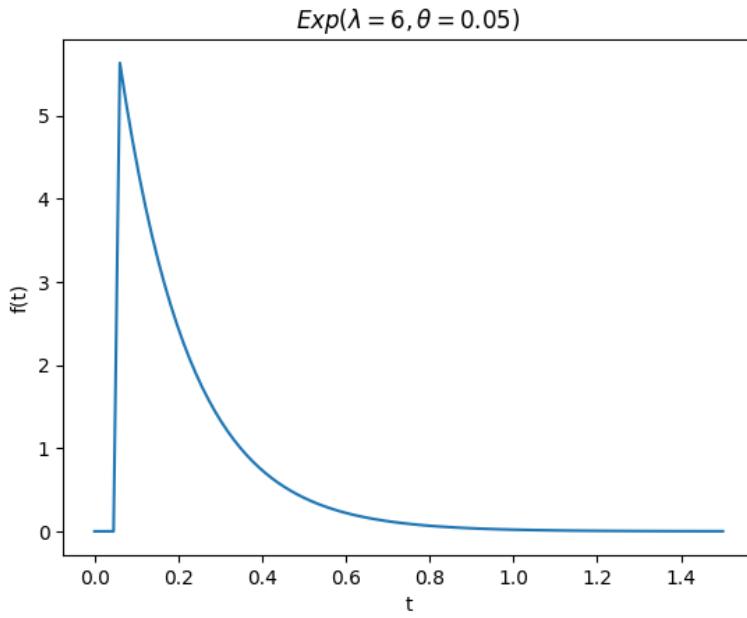


Рис. 82: Пороговое значение экспоненциального распределения, $Exp(\lambda = 6, \theta = 0.05)$

```

theta = 0.05

for t_period in time:
    print(expon.cdf(t_period, loc = theta, scale = b))

Output:
0.5034146962085905
0.9327944872602503
0.9966540345425288

```

Листинг 88: Вероятность T_1 через 10 минут, полчаса и час

Матожидание и дисперсия. Поскольку в общем случае для непрерывной случайной величины справедливо, что

$$\mathbb{E}[X] = \int x \cdot f(x) dx$$

для экспоненциального распределения найдем следующий несобственный интеграл

$$\mathbb{E}[t] = \int_0^{+\infty} t \cdot \lambda \exp(-\lambda t) dt = \lambda \int_0^{+\infty} t \cdot \exp(-\lambda t) dt$$

Найдем первообразную.

$$\int t \cdot \exp(-\lambda t) dt = \left(-\frac{1}{\lambda}x - \frac{1}{\lambda^2} \right) \exp(-\lambda x)$$

По обобщенной формуле Ньютона-Лейбница

$$\int_a^{+\infty} f(x) dx = \lim_{b \rightarrow +\infty} F(x) \Big|_a^b = \lim_{b \rightarrow +\infty} (F(b) - F(a))$$

получаем, что

$$\begin{aligned}\mathbb{E}[t] &= \lambda \left[\left(-\frac{1}{\lambda}t - \frac{1}{\lambda^2} \right) \exp(-\lambda t) \right]_0^{+\infty} \\ &= \lambda \left[\lim_{b \rightarrow \infty} \left(-\frac{1}{\lambda}b - \frac{1}{\lambda^2} \right) \exp(-\lambda b) - \left(-\frac{1}{\lambda} \cdot 0 - \frac{1}{\lambda^2} \right) \exp(-\lambda \cdot 0) \right] \\ &= \lambda \left[0 + \frac{1}{\lambda^2} \right] = \frac{1}{\lambda}\end{aligned}$$

Поскольку экспоненциальное распределение моделирует вероятность времени ожидания до первого события, логично, что его матожидание отражает среднее время, за которое происходит одно событие, то есть среднее время ожидания (параметр $\beta = \frac{1}{\lambda}$ при пороговом значении $\theta = 0$).

Приведем формулу дисперсии без доказательства

$$\mathbb{V}[t] = \frac{1}{\lambda^2}$$

Геометрическое распределение. Рассмотрим **геометрическое распределение** (geometric distribution) и его связь с экспоненциальным распределением. Это распределение дискретной случайной величины, которое показывает вероятность количества независимых испытаний Бернулли до первого «успеха».

Например, мы можем подбрасывать монету (по одному разу за испытание) с вероятностью выпадения решки равной p . Вероятность того, что решка выпадет после k испытаний и следует геометрическому распределению.

$$X_1 \sim Geom(p)$$

Ход испытаний и его результаты можно представить следующим образом (рисунок 83).

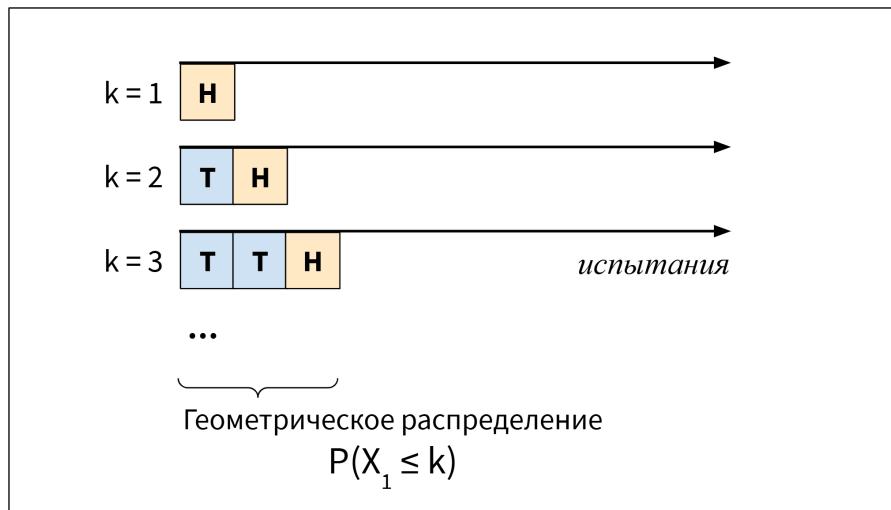


Рис. 83: Геометрический процесс

Приведем функцию вероятности (pmf).

$$P(X_1 = k) = (1 - p)^{k-1} p$$

Имеется в виду, что до первого успеха мы провели k испытаний Бернулли, из которых $k - 1$ закончились неудачей с вероятностью $1 - p$ и одно из них закончилось успехом с вероятностью p . Поскольку испытания независимы, их совместная вероятность есть просто их произведение.

Функция распределения (cdf) показывает, с какой вероятностью случайная величина X_1 примет значение меньшее или равное k , $P(X_1 \leq k)$.

Начнем выписывать значения функции распределения для $k \in \{1, 2, 3, \dots\}$

$$\begin{aligned} P(X_1 \leq 1) &= (1 - p)^{1-1}p = p \\ P(X_1 \leq 2) &= p(1 - p) + p \\ P(X_1 \leq 3) &= p(1 - p)^2 + p(1 - p) + p \\ &\dots \\ P(X_1 \leq k) &= p \left(\sum_{i=1}^k (1 - p)^{i-1} \right) \end{aligned}$$

Получается убывающая геометрическая прогрессия (отсюда и название распределения) с первым членом p и основанием $1 - p$, где $|1 - p| < 1$. Сумма бесконечной геометрической прогрессии, как и должно быть в данном случае, равна

$$\lim_{k \rightarrow \infty} \frac{p}{1 - (1 - p)} = \frac{p}{p} = 1$$

При этом сумма первых k членов прогрессии будет равна

$$P(X_1 \leq k) = \frac{p(1 - p)^k - 1}{(1 - p) - 1} = -((1 - p)^k - 1) = 1 - (1 - p)^k$$

Например, вероятность достичь первого успеха в не более чем пяти испытаниях при $p = \frac{1}{6}$ равна

$$P(X_1 \leq 5) = 1 - \left(1 - \frac{1}{6}\right)^5 \approx 0,60$$

Проверим с помощью графика (листинг 89 и рисунок 84) и вычислим интеграл (листинг 90).

```
from scipy.stats import geom

p = 1/6
k = np.arange(1, 25)

plt.plot(k, geom.pmf(k, p), 'bo', ms=8)
plt.vlines(k, 0, geom.pmf(k, p), colors='b', lw=5, alpha=0.5)
plt.title('Geometric distribution')
plt.xlabel('k')
plt.ylabel(r'$f(k | p = 1/6)$')
plt.show()
```

Листинг 89: Геометрическое распределение, $Geom(p = \frac{1}{6})$

```
geom.cdf(5, p)

Output: 0.5981224279835391
```

Листинг 90: Геометрическое распределение, $Geom(p = \frac{1}{6})$, $P(X_1 \leq 5)$

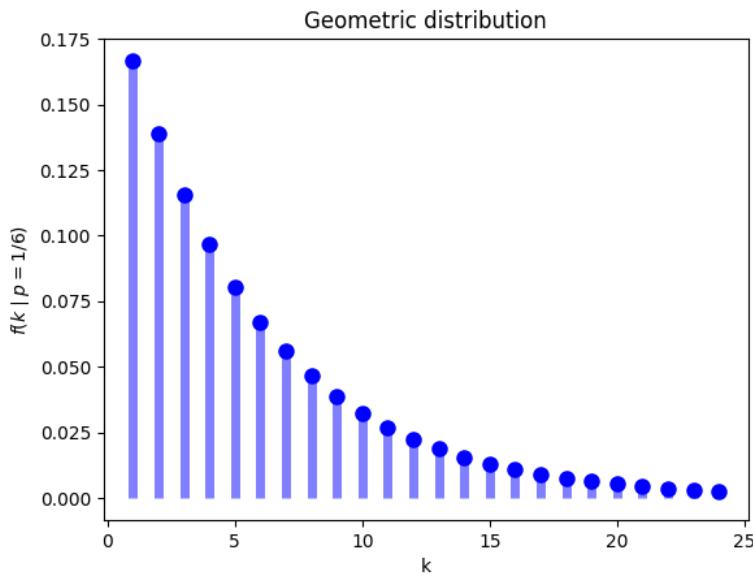


Рис. 84: Геометрическое распределение, $Geom(p = \frac{1}{6})$

Найдем матожидание.

$$\mathbb{E}[X] = \sum xP(X = x)$$

$$\mathbb{E}[k] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$$

Для простоты положим $a = p$ и $r = 1 - p$. Тогда, по формуле суммы бесконечной геометрической прогрессии

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}, |r| < 1$$

Найдем производную обеих частей тождества относительно r .

$$\frac{d}{dr} \left[\sum_{k=0}^{\infty} ar^k \right] = \frac{d}{dr} \left[\frac{a}{1-r} \right]$$

$$\sum_{k=0}^{\infty} akr^{k-1} = \frac{a}{(1-r)^2}$$

Обратим внимание, что при $k = 0$, $a0r^{0-1} = 0$, поэтому

$$\sum_{k=1}^{\infty} akr^{k-1} = \frac{a}{(1-r)^2}$$

$$\mathbb{E}[k] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = \frac{p}{(1-(1-p))^2} = \frac{p}{p^2} = \frac{1}{p}$$

Другими словами, если вероятность успеха в одном испытании равна, например, $p = \frac{1}{6}$, то для достижения первого успеха в среднем потребуется провести $\mathbb{E}[k] = 6$ испытаний (листинг 91).

```
geom.mean(p)
```

```
Output: 6.0
```

Листинг 91: Матожидание $Geom\left(p = \frac{1}{6}\right)$

Приведем формулу дисперсии без доказательства

$$\mathbb{V}[k] = \frac{1-p}{p^2}$$

На этом этапе становится довольно очевидно, что если рассматривать испытания Бернулли геометрического распределения как стремящиеся к бесконечности дискретные отрезки времени, то в результате мы получим экспоненциальное распределение (рисунок 85).

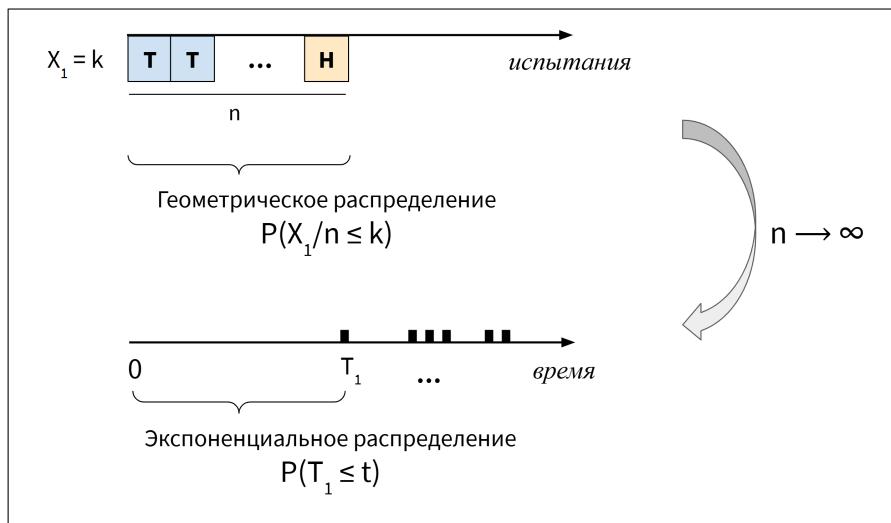


Рис. 85: Связь геометрического и экспоненциального распределений

Докажем это. Разделим каждую серию испытаний до первого успеха на n (см. рисунок 85). Тогда,

$$cdf(k) = P\left(\frac{X_1}{n} \leq k\right) = P(X_1 \leq nk)$$

Поскольку значение nk должно быть целым числом, применим функцию «пол» $\lfloor nk \rfloor$.

$$cdf(k) = P\left(\frac{X_1}{n} \leq k\right) = P(X_1 \leq \lfloor nk \rfloor) = 1 - (1-p)^{\lfloor nk \rfloor}$$

Положим $np = \lambda$ и $k = t$, тогда

$$\begin{aligned} \lim_{n \rightarrow \infty} [(1-p)^n] &= \lim_{n \rightarrow \infty} \left[\left(1 - \frac{np}{n}\right)^n \right] = e^{-np} = e^{-\lambda} \\ \lim_{n \rightarrow \infty} [(1-p)^{nt}] &= e^{-\lambda t} \end{aligned}$$

Теперь найдем $\lim_{n \rightarrow \infty} [(1-p)^{\lfloor nt \rfloor}]$. По определению функции «пол»

$$\begin{aligned} \lfloor nt \rfloor &\leq nt < \lfloor nt \rfloor + 1 \\ nt - 1 &< \lfloor nt \rfloor \leq nt \end{aligned}$$

Значит, при $n \rightarrow \infty$, получаем

$$\lim_{n \rightarrow \infty} [(1-p)^{\lfloor nt \rfloor}] = e^{-\lambda t}$$

т.е. функцию экспоненциального распределения.

Свойство отсутствия памяти. В геометрическом распределении мы производим независимые испытания Бернулли, а значит тот факт, что в первых a испытаниях не было ни одного «успеха», никак не влияет на вероятность успеха в последующих b испытаниях.

Такое свойство называется свойством **отсутствия памяти** (memoryless property).

$$P(X > a + b \mid X > a) = P(X > b)$$

Приведем пример. Предположим, что мы бросили монету $a = 5$ раз и решка не выпала ни разу (то есть не было ни одного «успеха»). В этом случае вероятность успеха после дополнительных $b = 10$ бросков будет такой же, как если бы мы начинали с нуля и оценивали вероятность успеха после десяти бросков.

$$P(X > 5 + 10 \mid X > 5) = P(X > 10)$$

Докажем это. Если $P(X \leq k) = 1 - (1 - p)^k$, то $P(X > k) = (1 - p)^k$. Тогда, по формуле совместной вероятности

$$P(X > a + b \mid X > a) = \frac{P(X > a + b, X > a)}{P(X > a)}$$

Запись $P(X > a + b, X > a)$ избыточна, поскольку, если $X > a + b$, то очевидно, что $X > a$. Тогда,

$$P(X > a + b \mid X > a) = \frac{P(X > a + b)}{P(X > a)} = \frac{(1 - p)^{a+b}}{(1 - p)^a} = (1 - p)^b = P(X > b)$$

Можно сказать, что доля вероятности $P(X > a + b)$ в $P(X > a)$ равна доле вероятности $P(X > b)$ в общем «объеме» распределения.

$$\frac{P(X > a + b)}{P(X > a)} = \frac{P(X > b)}{1}$$

Приведем иллюстрацию (рисунки 86 и 87), а также убедимся в правильности выводов с помощью Питона (листинг 92).

```
a, b = 5, 10

# P(X > a + b)
numerator = 1-geom.cdf(a+b, p)

# P(X > a)
denominator = 1-geom.cdf(a, p)

numerator/denominator

Output: 0.16150558288984576

# P(X > b)
1-geom.cdf(b, p)

Output: 0.16150558288984573
```

Листинг 92: Свойство отсутствия памяти. Пример

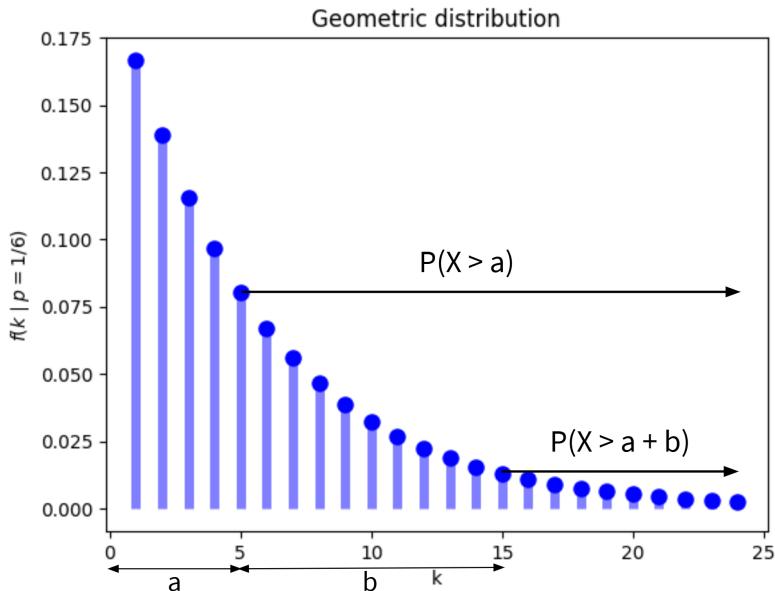


Рис. 86: Свойство отсутствия памяти

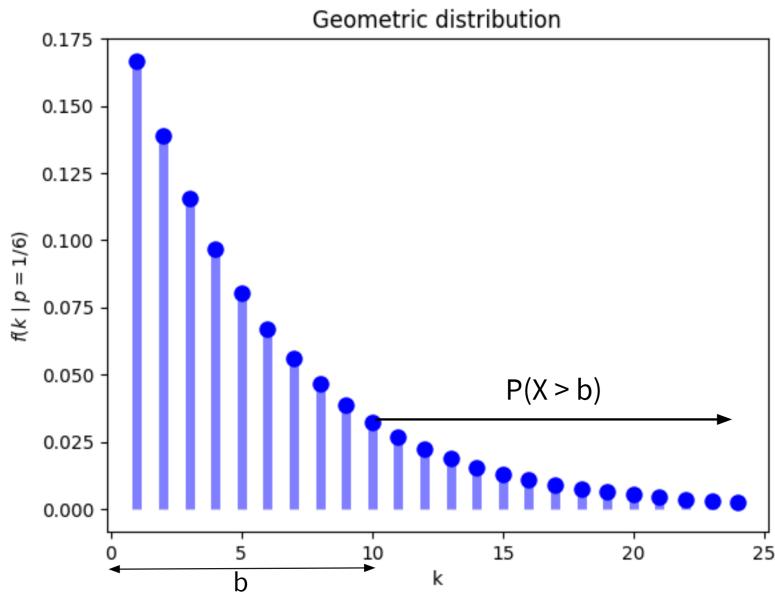


Рис. 87: Свойство отсутствия памяти. Продолжение

Экспоненциальное распределение также обладает свойством отсутствия памяти. Приведем аналогичное доказательство.

Поскольку $P(T > t) = e^{-\lambda t}$, то

$$P(T > a + b \mid T > a) = \frac{P(T > a + b)}{P(T > a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P(T > b)$$

В случае моделирования очереди свойство отсутствия памяти указывает на то, что вероятность прихода первого покупателя после ожидания вначале в течение $a = 3$ минут, а затем еще $b = 9$ минут равна вероятности его прихода после ожидания в течение $b = 9$ минут (листинг 93).

```

lambda_ = 6

# 3 min, 9 min
a, b = 0.05, 0.15

# P(X > a + b) / P(X > a)
(1-expon.cdf(a+b, scale = 1/lambda_))/(1-expon.cdf(a, scale = 1/lambda_))

Output: 0.406569659740599

# P(X > b)
1-expon.cdf(b, scale = 1/lambda_)

Output: 0.4065696597405991

```

Листинг 93: Свойство отсутствия памяти экспоненциального распределения

Заметим, что геометрическое и экспоненциальное распределения — единственные распределения, обладающие свойством отсутствия памяти.

Распределение Вейбулла. В случае моделирования вероятности ожидания первого покупателя с помощью экспоненциального распределения свойство отсутствия памяти не вызывает сомнений, поскольку логично предположить, что поведение одного покупателя никак не зависит от поведения следующего.

При этом, если моделируется вероятность времени до первого отказа токарного станка, то очевидно, что если станок уже проработал шесть лет, то вероятность отказа в последующие три года никак не может быть равна вероятности отказа в первые три года эксплуатации нового оборудования.

В этом случае говорят об изменяющейся интенсивности случайного процесса (non-homogenous Poisson process), и моделировать время ожидания до первого события можно с помощью **распределения Вейбулла** (Weibull distribution).

$$pdf(t; \lambda, c) = \frac{c}{\lambda} \left(\frac{t}{\lambda}\right)^{c-1} e^{-\left(\frac{t}{\lambda}\right)^c}, \quad t \geq 0$$

$$cdf(t; \lambda, c) = 1 - e^{-\left(\frac{t}{\lambda}\right)^c}, \quad t \geq 0$$

- при $c < 1$ интенсивность процесса снижается;
- при $c = 1$ интенсивность остается неизменной и распределение Вейбулла совпадает с экспоненциальным при одинаковом параметре λ ;
- при $c > 1$ интенсивность возрастает.

Посмотрим на график (листинг 94 и рисунок 88).

```

%%capture --no-display

from scipy.stats import weibull_min

c1, c2, c3 = 0.8, 1, 1.2
lambda_ = 6

# scale = 1/lambda
t = np.linspace(0, 1.2, 100)
plt.plot(t, weibull_min.pdf(t, c1, scale=1/lambda_), label = 'c = {}'.
          format(c1))
plt.plot(t, weibull_min.pdf(t, c2, scale=1/lambda_), label = 'c = {}'.
          format(c2))
plt.plot(t, weibull_min.pdf(t, c3, scale=1/lambda_), label = 'c = {}'.
          format(c3))
plt.title(r'Weibull distribution, $WB(c, \lambda = 6)$')
plt.xlabel('t')
plt.ylabel('f(t)')
plt.legend()
plt.show()

```

Листинг 94: Распределение Вейбулла

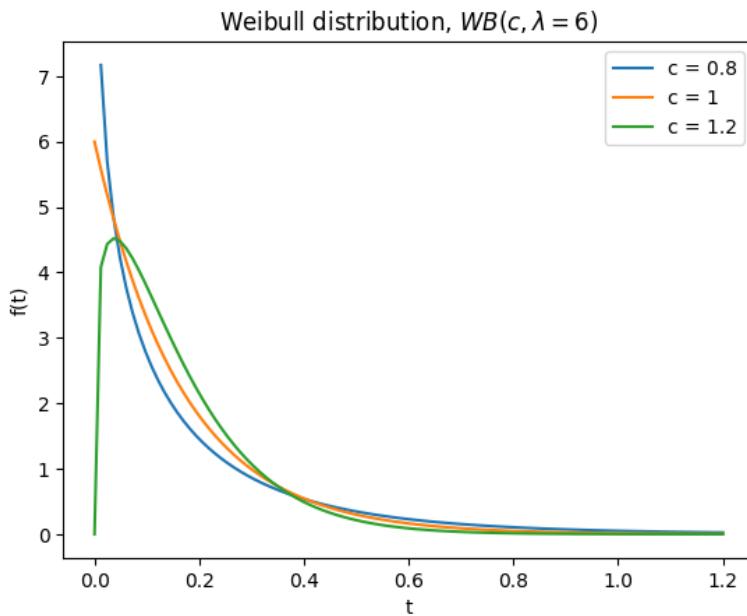


Рис. 88: Распределение Вейбулла

Поскольку распределение Вейбулла часто используют для моделирования вероятности времени до первого отказа оборудования(time-to-failure), интенсивность процесса называют **интенсивностью отказов** (failure rate). Посмотрим на изменение интенсивности процесса при различных значениях c .

Рассчитать интенсивность в конкретный момент времени t в распределении Вейбулла можно по формуле

$$f(t; c, \beta) = \beta c (\beta t)^{c-1}, \text{ где } \beta = \frac{1}{\lambda}$$

Приведем расчет с помощью Питона (листинг 95).

```

def failure_rate(t, c, shape):
    return (shape * c) * (shape * t) ** (c-1)

time = [0.1, 0.5, 1]

# c < 1
for t in time:
    print(failure_rate(t, c1, 1 / lambda_))

Output:
0.3023910873688072
0.21916691060229673
0.1907958774807007

# c = 1
for t in time:
    print(failure_rate(t, c2, 1 / lambda_))

Output:
0.16666666666666666
0.16666666666666666
0.16666666666666666

# c > 1
for t in time:
    print(failure_rate(t, c3, 1 / lambda_))

Output:
0.08818602062217205
0.12167286837864116
0.13976542375431586

```

Листинг 95: Расчет интенсивности

Как и ожидалось, интенсивность при $c = 1$ неизменна и соответствует одному отказу каждые $0,1\bar{6} = \frac{1}{6}$ часа или каждые 10 минут. При $c < 1$ и $c > 1$ интенсивность убывает и возрастает соответственно.

Вернемся к изучению процесса Пуассона с неизменной интенсивностью (homogenous Poisson process) и наконец определимся с сопряженным распределением Пуассона априорным распределением и научимся оценивать апостериорную вероятность.

7.4 Время ожидания k-ого события

В примере с подходящими к кассе покупателями, если мы сами стоим в этой очереди, скажем, третьими, то логично оценивать вероятность именно нашего времени ожидания, то есть T_3 , а не первого человека в очереди T_1 . Посмотрим, как это можно сделать.

7.4.1 Распределение Эрланга

Вспомним, чтобы оценить вероятность времени ожидания T_1 с помощью распределения Пуассона, мы нашли вероятность того, что за время t не произойдет ни одного события. Другими словами,

$$P(T_1 > t) = P(X = 0 \mid \lambda t) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!}$$

Если же мы хотим оценить вероятность ожидания T_3 , то нам нужно вначале оценить вероятность того, что произойдет $k = 0, 1$ или 2 события. Поскольку для взаимоисключающих событий $P(A \cup B) = P(A) + P(B)$, то

$$\begin{aligned} P(T_3 > t) &= P(X = 0 \mid \lambda t) + P(X = 1 \mid \lambda t) + P(X = 2 \mid \lambda t) \\ &= \frac{(\lambda t)^0 e^{-\lambda t}}{0!} + \frac{(\lambda t)^1 e^{-\lambda t}}{1!} + \frac{(\lambda t)^2 e^{-\lambda t}}{2!} \end{aligned}$$

Тогда вероятность ожидания трех событий равна

$$P(T_3 \leq t) = 1 - \left(\frac{(\lambda t)^0 e^{-\lambda t}}{0!} + \frac{(\lambda t)^1 e^{-\lambda t}}{1!} + \frac{(\lambda t)^2 e^{-\lambda t}}{2!} \right)$$

В общем виде нам нужно найти

$$P(T_k \leq t) = 1 - P(T_k > t) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda t)^i e^{-\lambda t}}{i!},$$

где $k \in \{1, 2, 3, \dots\}$ и $\lambda \in (0, \infty)$. Такая функция распределения вероятностей описывает **распределение Эрланга** (Erlang distribution).

Оценим вероятность ожидания не более 15 минут (0,25 часа) до наступления трех событий при $\lambda = 6$.

$$cdf(0,25 \mid 3, 6) = 1 - \sum_{i=0}^{3-1} \frac{(6 \cdot 0,25)^i e^{-6 \cdot 0,25}}{i!} \approx 0,19$$

Проверим с помощью Питона (листинг 96).

```
from scipy.stats import erlang

t, k, lambda_ = 0.25, 3, 6

erlang.cdf(t, k, scale=1/lambda_)

Output: 0.19115316946194183
```

Листинг 96: Распределение Эрланга. Вероятность ожидания не более 15 минут

Если вы спешите, и у вас есть не более 15 минут на ожидание в очереди, при такой вероятности лучше прийти в другой раз.

Для того чтобы найти плотность вероятности, дифференцируем функцию распределения относительно t .

$$pdf(t \mid k, \lambda) = \frac{d}{dt} \left(1 - \sum_{i=0}^{k-1} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \right)$$

Вынесем первое слагаемое при $k = 0$ за знак суммы.

$$= \frac{d}{dt} \left(1 - e^{-\lambda t} - \sum_{i=1}^{k-1} \frac{1}{i!} [(\lambda t)^i e^{-\lambda t}] \right)$$

Воспользуемся правилами производной степенной функции, произведения, а также производной экспоненциальной функции $(e^x)' = e^x$.

$$= \lambda e^{-\lambda t} - \sum_{i=1}^{k-1} \frac{1}{i!} [(\lambda t)^i \cdot (-\lambda e^{-\lambda t}) + e^{-\lambda t} \cdot i(\lambda t)^{i-1} \cdot \lambda]$$

Вынесем $-\lambda e^{-\lambda t}$ за скобки, а затем из-под знака суммы. Одновременно разделим скобку на $i!$

$$= \lambda e^{-\lambda t} + \lambda e^{-\lambda t} \left[\sum_{i=1}^{k-1} \left(\frac{(\lambda t)^i}{i!} - \frac{\lambda(\lambda t)^{i-1}}{(i-1)!} \right) \right]$$

Выпишем некоторые слагаемые суммы от $i = 1$ до $i = k - 1$.

$$= \lambda e^{-\lambda t} + \lambda e^{-\lambda t} \cdot \left[(\lambda t - 1) + \left(\frac{(\lambda t)^2}{2!} - \lambda t \right) + \dots + \left(\frac{(\lambda t)^{k-1}}{(k-1)!} - \frac{(\lambda t)^{k-2}}{(k-2)!} \right) \right]$$

Обратим внимание, что большинство слагаемых сократится.

$$= \lambda e^{-\lambda t} + \lambda e^{-\lambda t} \left[-1 + \frac{(\lambda t)^{k-1}}{(k-1)!} \right]$$

Упростив выражение, получим

$$= \lambda e^{-\lambda t} - \lambda e^{-\lambda t} + \frac{\lambda e^{-\lambda t} (\lambda t)^{k-1}}{(k-1)!} = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}$$

7.4.2 Гамма-распределение

Изменим нотацию в распределении Эрланга. Пусть $k = \alpha$, $\lambda = \beta$, тогда

$$pdf(t | \alpha, \beta) = \underbrace{(\beta^\alpha / (\alpha - 1)!) \cdot}_{\text{normalizing factor}} \underbrace{(t^{\alpha-1} e^{-\beta t})}_{\text{kernel}}$$

Примечание. Обратите внимание, как мы записали функцию плотности:

- та ее часть, которая не содержит переменной t , называется **нормализующим коэффициентом** (normalizing factor) и зачастую, как мы уже видели, опускается;
- вторая часть, содержащая t , называется **ядром** (kernel) распределения.

При этом параметры, в частности, α и β , могут присутствовать в обеих частях.

Теперь рассмотрим ситуацию, при которой α может принимать положительные вещественные значения $\alpha \in \mathbb{R}_{>0}$, а не только целые положительные числа. В этом случае $(\alpha - 1)! = \Gamma(\alpha)$. В частности,

$$pdf(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}$$

Такая функция плотности задает **гамма-распределение** (gamma distribution) с параметрами α и β .

$$t \sim Gamma(\alpha, \beta)$$

Матожидание и дисперсия. Найдем матожидание гамма-распределения. Поскольку,

$$\mathbb{E}[X] = \int x \cdot f(x) dx$$

тогда, с учетом функции плотности вероятности, получаем

$$\begin{aligned}\mathbb{E}[t] &= \int_0^\infty t \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t) dt \\ &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} t^{(\alpha+1)-1} \exp(-\beta t) dt \\ &= \int_0^\infty \frac{1}{\beta} \cdot \frac{\beta^{\alpha+1}}{\Gamma(\alpha)} t^{(\alpha+1)-1} \exp(-\beta t) dt\end{aligned}$$

Поскольку $\Gamma(z+1) = z\Gamma(z)$ и $\Gamma(z) = \Gamma(z+1) \cdot z^{-1}$, то

$$\begin{aligned}&= \int_0^\infty \frac{\alpha}{\beta} \cdot \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} t^{(\alpha+1)-1} \exp(-\beta t) dt \\ &= \frac{\alpha}{\beta} \int_0^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} t^{(\alpha+1)-1} \exp(-\beta t) dt \\ &= \frac{\alpha}{\beta} \int_0^\infty \text{Gamma}(t | \alpha+1, \beta) dt\end{aligned}$$

В данном случае $\int_0^\infty \text{Gamma}(t | \alpha+1, \beta) dt = 1$. Как следствие,

$$\mathbb{E}[t] = \frac{\alpha}{\beta}$$

Приведем еще одно доказательство.

$$\begin{aligned}\mathbb{E}[t] &= \int_0^\infty t \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t) dt \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty t^{(\alpha+1)-1} \exp(-\beta t) dt\end{aligned}$$

Если положить $x = \beta t$, то

$$\begin{aligned}&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{x}{\beta}\right)^{(\alpha+1)-1} e^{-x} \frac{dx}{\beta} \\ &= \frac{\beta^\alpha}{\beta^{\alpha+1} \Gamma(\alpha)} \int_0^\infty x^{(\alpha+1)-1} e^{-x} dx \\ &= \frac{1}{\beta \Gamma(\alpha)} \int_0^\infty x^{(\alpha+1)-1} e^{-x} dx\end{aligned}$$

По определению гамма-функции $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$

$$= \frac{\Gamma(\alpha+1)}{\beta \Gamma(\alpha)} = \frac{\alpha \Gamma(\alpha)}{\beta \Gamma(\alpha)} = \frac{\alpha}{\beta}$$

Приведем формулу дисперсии без доказательства.

$$\mathbb{V}[t] = \frac{\alpha}{\beta^2}$$

Функция распределения. Выведем функцию распределения вероятности cdf . Напомню, нам нужно найти накопленную вероятность таких возможных значений функции плотности $pdf(z)$, которые меньше или равны t . Другими словами,

$$\begin{aligned} cdf(t) &= \int_0^t Gamma(z | \alpha, \beta) dz \\ &= \int_0^t \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z} dz \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^t z^{\alpha-1} e^{-\beta z} dz \end{aligned}$$

Положим $x = \beta z$ и $z = \frac{x}{\beta}$. Тогда,

$$\begin{aligned} cdf(t) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_{\beta 0}^{\beta t} \left(\frac{x}{\beta} \right)^{\alpha-1} \exp \left[-\beta \left(\frac{x}{\beta} \right) \right] d \left(\frac{x}{\beta} \right) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{\beta^{\alpha-1}} \cdot \frac{1}{\beta} \int_0^{\beta t} x^{\alpha-1} \exp[-x] dx \\ &= \frac{1}{\Gamma(\alpha)} \int_0^{\beta t} x^{\alpha-1} \exp[-x] dx \end{aligned}$$

Интеграл определяет нижнюю неполную гамма-функцию (lower incomplete gamma function) γ с переменным верхним пределом βt .

$$\gamma(\alpha, \beta t) = \int_0^{\beta t} x^{\alpha-1} \exp[-x] dx$$

Тогда функция распределения, что логично, представляет собой отношение значения нижней неполной гамма-функции и значения (полней) гамма-функции.

$$cdf(t) = \frac{\gamma(\alpha, \beta t)}{\Gamma(\alpha)}$$

Теперь рассмотрим параметры α и β более подробно.

Параметры распределения. Гамма-распределение оценивает время ожидания до наступления α -ого события процесса Пуассона, происходящих с интенсивностью (средним временем между событиями) β .

Зафиксируем $\alpha = 3$ и посмотрим, как будет меняться β (листинг 97 и рисунок 89).

```

from scipy.stats import gamma

t = np.linspace(0, 2.5, 100)
a = 3
b1, b2, b3 = 4, 6, 8

plt.plot(t, gamma.pdf(t, a, scale = 1/b1), label = r'$\beta_1$ = {}'.
          format(b1))
plt.plot(t, gamma.pdf(t, a, scale = 1/b2), label = r'$\beta_2$ = {}'.
          format(b2))
plt.plot(t, gamma.pdf(t, a, scale = 1/b3), label = r'$\beta_3$ = {}'.
          format(b3))
plt.title(r'Gamma distribution, $Gamma(\alpha = 3, \beta)$')
plt.xlabel('t')
plt.ylabel('f(t)')
plt.legend()
plt.show()

```

Листинг 97: $Gamma(\alpha = 3, \beta)$

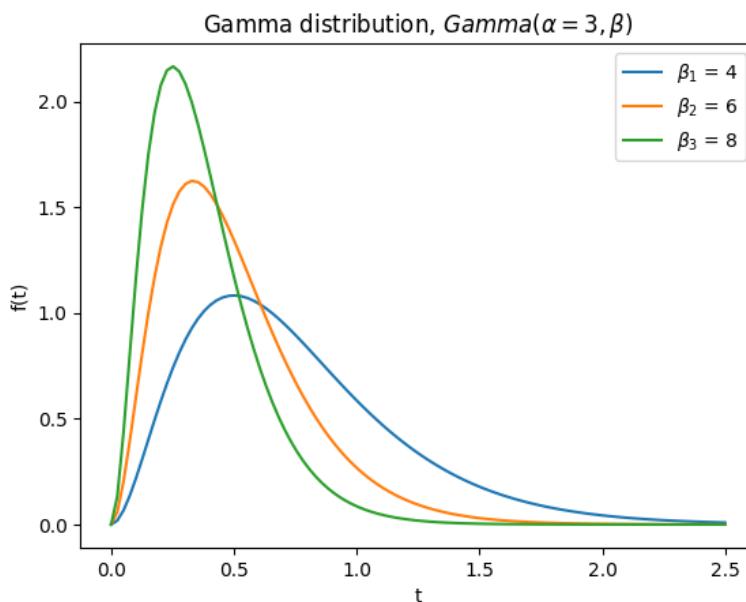


Рис. 89: $Gamma(\alpha = 3, \beta)$

На графике видно, что с ростом интенсивности β время ожидания уменьшается. Аналогично, можно сказать, что с увеличением среднего времени между событиями $\frac{1}{\beta}$ время ожидания увеличивается (листинг 98).

```

for b in [b3, b2, b1]:
    print(gamma.cdf(1, a, scale = 1/b))

Output:
0.986246032255997
0.938031195583341
0.7618966944464556

```

Листинг 98: $Gamma(\alpha = 3, \beta)$

Теперь зафиксируем интенсивность $\beta = 3$ и посмотрим, как меняется вероятность времени ожидания с изменением α (листинг 99 и рисунок 90).

```
a1, a2, a3 = 1, 2, 3
b = 3

plt.plot(t, gamma.pdf(t, a1, scale = 1/b), label = r'$\alpha_1$ = {}'.
          format(a1))
plt.plot(t, gamma.pdf(t, a2, scale = 1/b), label = r'$\alpha_2$ = {}'.
          format(a2))
plt.plot(t, gamma.pdf(t, a3, scale = 1/b), label = r'$\alpha_3$ = {}'.
          format(a3))
plt.title(r'$\Gamma(\alpha, \beta = 3)$')
plt.xlabel('t')
plt.ylabel('f(t)')
plt.legend()
plt.show()
```

Листинг 99: $\Gamma(\alpha, \beta = 3)$

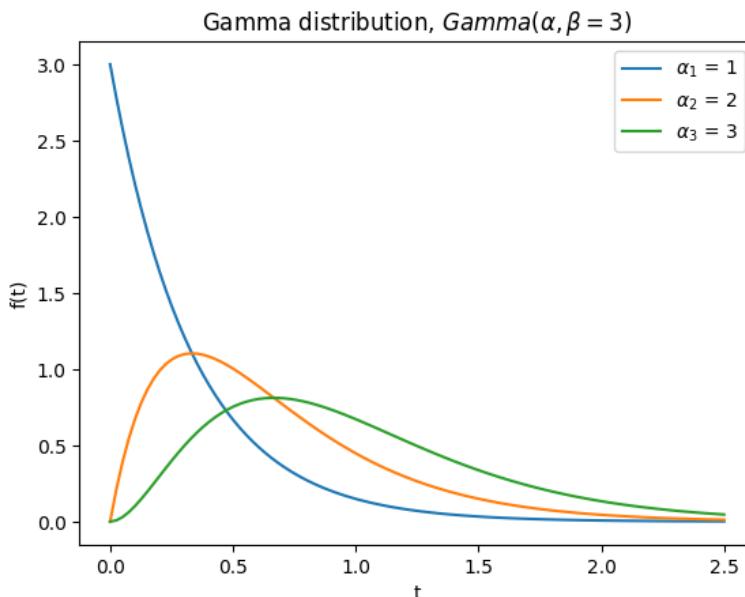


Рис. 90: $\Gamma(\alpha, \beta = 3)$

При одинаковой интенсивности с увеличением α время ожидания увеличивается (листинг 100).

```
for a in [a1, a2, a3]:
    print(gamma.cdf(1, a, scale = 1/b))

Output:
0.950212931632136
0.8008517265285442
0.5768099188731566
```

Листинг 100: $\Gamma(\alpha, \beta = 3)$

Другими словами, наступления третьего события придется ждать дольше, чем наступле-

ния первого. Перейдем к вопросу сопряженности распределений.

7.5 Сопряженность распределений

Как мы видим, гамма-распределение — это непрерывное распределение с носителем $(0, \infty)$ и параметрами α и β . Его удобно использовать в качестве априорного распределения для параметра λ .

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$

$$P(\lambda | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Кроме этого, интуитивно, гамма-распределение и распределение Пуассона (с учетом всего вышеизложенного) должны обладать свойством сопряженности. Докажем это формально.

Доказательство сопряженности. Пусть k_i — количество n событий, произошедших за определенный период времени. Например, количество покупателей, подошедших к кассе в определенный i час. Тогда правдоподобие распределено

$$k_i \sim \text{Pois}(\lambda)$$

$$\mathcal{L}(k_i | \lambda) = \frac{\lambda^{k_i} e^{-\lambda}}{k_i!}$$

По формуле Байеса

$$P(\lambda | \mathbf{k}) = \frac{\mathcal{L}(\mathbf{k} | \lambda) \cdot P(\lambda)}{P(\mathbf{k})} \propto \mathcal{L}(\mathbf{k} | \lambda) \cdot P(\lambda),$$

где

$$\mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$

С учетом независимости k_i наблюдений, их совместное правдоподобие будет равно

$$\mathcal{L}(\mathbf{k} | \lambda) = \prod_{i=1}^n \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \frac{\lambda^{k_1+k_2+k_3+\dots+k_n} \cdot e^{-n\lambda}}{\prod_{i=1}^n k_i!} = \frac{\lambda^{\sum_{i=1}^n k_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n k_i!}$$

Знаменатель не содержит λ , поэтому

$$\mathcal{L}(\mathbf{k} | \lambda) \propto \lambda^{\sum_{i=1}^n k_i} \cdot e^{-n\lambda}$$

Так как $\bar{k} = \frac{\sum_{i=1}^n k_i}{n}$, то $\sum_{i=1}^n k_i = n\bar{k}$, а значит

$$\mathcal{L}(\mathbf{k} | \lambda) \propto \lambda^{n\bar{k}} \cdot e^{-n\lambda}$$

Априорное распределение можно записать и так

$$P(\lambda | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \propto \lambda^{\alpha-1} e^{-\beta\lambda}$$

Тогда по формуле Байеса

$$P(\lambda | \mathbf{k}) \propto \mathcal{L}(\mathbf{k} | \lambda) \cdot P(\lambda) = \lambda^{n\bar{k}} \cdot e^{-n\lambda} \cdot \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{n\bar{k}+\alpha-1} e^{-(\beta+n)\lambda}$$

Добавим константу.

$$P(\lambda | \mathbf{k}) = \frac{(\beta + n)^{n\bar{k}+\alpha}}{\Gamma(n\bar{k} + \alpha)} \lambda^{n\bar{k}+\alpha-1} e^{-(\beta+n)\lambda} \sim Gamma(n\bar{k} + \alpha, \beta + n)$$

Обратим внимание на важный момент. В данном случае, априорное гамма-распределение моделирует наше изначальное представление о параметре λ , а не вероятность ожидания t до α -ого события в процессе Пуассона с интенсивностью β .

Пример. Закрепим это понимание на несложном примере.

Предположим, что мы хотим смоделировать распределение λ интенсивности звонков в колл-центр. Изначально мы считаем, что поступает около пяти звонков в час, однако разумно предположить, что их может быть от 2 до 7.

Подберем такие параметры α и β , чтобы они отражали наше априорное представление о распределении параметра λ . Среднее такого распределения должно быть равно

$$\mathbb{E}[\lambda] = \frac{\alpha}{\beta} = 5$$

При этом хотелось бы, чтобы большая часть значений находилась в пределах от 2 до 7. Подобным априорным убеждениям соответствует, например, распределение

$$\lambda \sim Gamma(10, 2)$$

Кроме этого мы получили следующие данные: в один и тот же час (например, с 14 до 15) в течение $n = 4$ дней поступило $k_1 = 5, k_2 = 3, k_3 = 3, k_4 = 1$ звонков. Таким образом, в среднем поступало 3 звонка в день (листинг 101).

```
a, b = 10, 2

data = np.array([5, 3, 3, 1])
mean_k, n = data.mean(), len(data)
mean_k, n

Output: (3.0, 4)
```

Листинг 101: Параметры гамма-распределения и собранные данные

Напишем функцию правдоподобия для отображения на графике (листинг 102).

$$\mathcal{L}(\mathbf{k} | \lambda) = \frac{\lambda^{\sum_{i=1}^n k_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n k_i!}$$

```

from math import factorial

def pois_likelihood(lambda_, data, scaling):
    sum_k = np.sum(data)
    n = len(data)

    numerator = lambda_ ** sum_k * np.exp(-n * lambda_)

    denominator = 1
    for k in data:
        denominator *= factorial(k)

    return (numerator/denominator) * scaling

```

Листинг 102: Функция правдоподобия

Примечание. Коэффициент масштабирования scaling позволяет вывести правдоподобие на одном графике с априорной и апостериорной вероятностью.

Найдем апостериорное распределение.

$$P(\lambda | \mathbf{k}) \sim \text{Gamma}(n\bar{k} + \alpha, \beta + n) = \text{Gamma}(4 \cdot 3 + 10, 2 + 4)$$

Построим графики (листинг 103 и рисунок 91).

```

lambda_ = np.linspace(0, 12, 100)
plt.plot(lambda_, pois_likelihood(lambda_, data, 500),
         label = '(scaled) likelihood')
plt.plot(lambda_, gamma.pdf(lambda_, a, scale = 1/b),
         label = 'prior')
plt.plot(lambda_, gamma.pdf(lambda_, (mean_k * n) + a, scale = 1/(b + n)),
         label = 'posterior')

plt.title(r'Gamma-Poisson model of $\lambda$')
plt.xlabel(r'$\lambda$')
plt.ylabel(r'$f(\lambda)$')
plt.legend()
plt.show()

```

Листинг 103: Модель сопряженных гамма-распределения и распределения Пуассона

Найдем среднее арифметическое и дисперсию априорного и апостериорного распределений (листинг 104).

```

gamma.stats(a, scale = 1/b, moments = 'mv')

Output: (5.0, 2.5)

gamma.stats((mean_k * n) + a, scale = 1/(b + n), moments = 'mv')

Output: (3.6666666666666665, 0.6111111111111111)

```

Листинг 104: Среднее арифметическое и дисперсия распределений

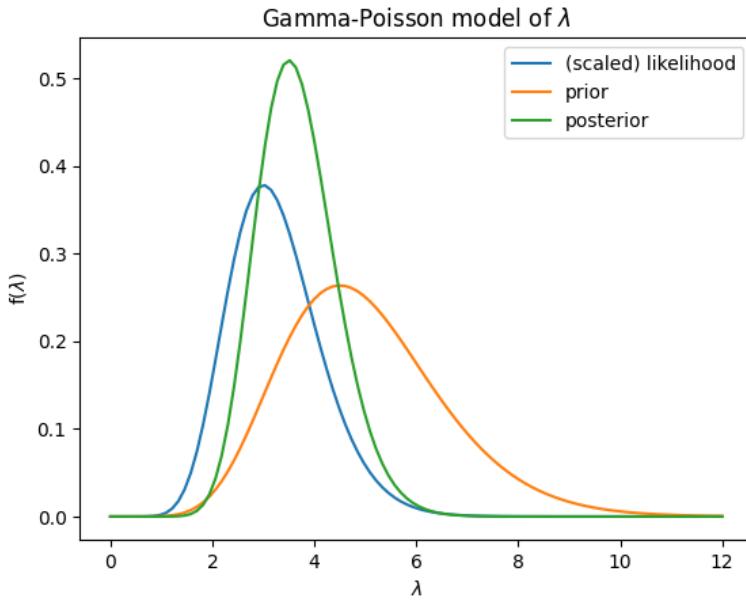


Рис. 91: Модель сопряженных гамма-распределения и распределения Пуассона

На основе графика и расчетов видно, что матожидание приблизилось к среднему значению полученных данных, а дисперсия снизилась, что отражает большую уверенность в распределении λ .

7.6 Прогнозные распределения

Прогнозное априорное распределение. Выведем прогнозное априорное распределение данных $P(X)$, то есть распределение данных X до их получения, усредненное по всем возможным значениям λ . Вспомним, что

$$k \sim Pois(\lambda)$$

Тогда,

$$\begin{aligned} P(X) &= \int_0^\infty P(X, \lambda) d\lambda = \int_0^\infty P(X | \lambda) \cdot P(\lambda) d\lambda \\ &= \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^\alpha}{k! \cdot \Gamma(\alpha)} \int_0^\infty \underbrace{\lambda^{k+\alpha-1} e^{-(\beta+1)\lambda}}_{\text{Gamma}(k+a, \beta+1)} d\lambda \end{aligned}$$

Обратим внимание, что подынтегральное выражение является ядром гамма-распределения. Дополним ядро константой, чтобы превратить его в полноценное распределение.

$$= \frac{\beta^\alpha}{k! \cdot \Gamma(\alpha)} \cdot \frac{\Gamma(k + \alpha)}{(\beta + 1)^{k+\alpha}} \int_0^\infty \frac{(\beta + 1)^{k+\alpha}}{\Gamma(k + \alpha)} \lambda^{k+\alpha-1} e^{-(\beta+1)\lambda} d\lambda$$

Интеграл равен единице. Как следствие,

$$= \frac{\beta^\alpha}{(\beta + 1)^{k+\alpha}} \cdot \frac{\Gamma(k + \alpha)}{k! \cdot \Gamma(\alpha)} = \left(\frac{\beta}{\beta + 1} \right)^\alpha \left(\frac{1}{\beta + 1} \right)^k \left(\frac{(k + \alpha - 1)!}{k!(\alpha - 1)!} \right)$$

Третий множитель представляет собой число сочетаний с повторениями $\binom{k+n-1}{k}$. Тогда,

$$\binom{k+\alpha-1}{k} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^k$$

задает **отрицательное биномиальное распределение** (negative binomial distribution) с параметрами α и $\frac{\beta}{\beta+1}$.

$$P(X) \sim NB(\alpha, \frac{\beta}{\beta+1}),$$

где β — шансы на успех в испытании (success odds, o_s). И действительно, если положить p — вероятность успеха и q — неудачи, то

$$o_s = \frac{p}{q} = \frac{p}{p-1},$$

откуда вероятность успеха $p = \frac{o_s}{o_s+1} = \frac{\beta}{\beta+1}$. Тогда отрицательное биномиальное распределение можно записать с более привычной параметризацией

$$P(X) \sim NB(r, p)$$

$$pmf(k) = \binom{k+r-1}{k} p^r (1-p)^k$$

Приведем еще одно доказательство. Из теоремы Байеса следует, что

$$P(\lambda | X) = \frac{P(X | \lambda) \cdot P(\lambda)}{P(X)}$$

$$P(X) = \frac{P(X | \lambda) \cdot P(\lambda)}{P(\lambda | X)}$$

Так как мы знаем априорное распределение и правдоподобие, а также уже нашли апостериорное распределение, то для одного наблюдения $n = 1$ получим

$$P(X) = \frac{\frac{\lambda^k \cdot e^{-\lambda}}{k!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}}{\frac{(\beta+1)^{k+\alpha}}{\Gamma(k+\alpha)} \lambda^{k+\alpha-1} e^{-(\beta+1)\lambda}}$$

Некоторые множители можно сократить.

$$= \frac{\cancel{\lambda^k \cdot e^{-\lambda}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cancel{\lambda^{\alpha-1} e^{-\beta\lambda}}}{\frac{(\beta+1)^{k+\alpha}}{\Gamma(k+\alpha)} \cancel{\lambda^{k+\alpha-1} e^{-(\beta+1)\lambda}}}$$

Обратим внимание, что сокращаются те множители, в которых есть λ , то есть параметр, относительно которого мы изначально интегрируем знаменатель формулы Байеса.

$$\begin{aligned} &= \frac{\beta^\alpha}{(\beta+1)^{k+\alpha}} \cdot \frac{\Gamma(k+\alpha)}{k! \cdot \Gamma(\alpha)} \\ &= \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^k \binom{k+\alpha-1}{k} \end{aligned}$$

Мы снова приходим к отрицательному биномиальному распределению.

Прогнозное апостериорное распределение. Найдем прогнозное апостериорное распределение $P(X' | \mathbf{X})$, то есть распределение нового $n = 1$ наблюдения X' после получения вектора данных \mathbf{X} . По формуле Байеса,

$$P(\lambda | X', \mathbf{X}) = \frac{P(X' | \lambda, \mathbf{X}) \cdot P(\lambda | \mathbf{X})}{P(X' | \mathbf{X})}$$

Еще раз заметим, что правдоподобие новых данных X' зависит только от параметра λ , а не от \mathbf{X} . Значит,

$$P(X' | \lambda, \mathbf{X}) = P(X' | \lambda)$$

Кроме того, обратим внимание, что $P(\lambda | \mathbf{X})$ — апостериорное гамма-распределение.

$$P(\lambda | \mathbf{X}) \sim \text{Gamma}(n\bar{k} + \alpha, n + \beta)$$

Тогда,

$$P(X' | \mathbf{X}) = \frac{P(X' | \lambda) \cdot P(\lambda | \mathbf{X})}{P(\lambda | X', \mathbf{X})}$$

Знаменатель $P(\lambda | X', \mathbf{X})$ при этом является новым апостериорным распределением, которое по свойству сопряженности следует гамма-распределению

$$P(\lambda | X', \mathbf{X}) \sim \text{Gamma}(n\bar{k} + k' + \alpha, n + 1 + \beta)$$

Другими словами, новое апостериорное распределение обновилось за счет $n = 1$ наблюдения k' .

$$\begin{aligned} P(X' | \mathbf{X}) &= \frac{\frac{\lambda^{k'} e^{-\lambda}}{k'!} \cdot \frac{(n+\beta)^{n\bar{k}+\alpha}}{\Gamma(n\bar{k}+\alpha)} \lambda^{n\bar{k}+\alpha-1} e^{-(n+\beta)\lambda}}{\frac{(n+1+\beta)^{n\bar{x}+k'+\alpha}}{\Gamma(n\bar{x}+k'+\alpha)} \lambda^{n\bar{x}+k'+\alpha-1} e^{-(n+1+\beta)\lambda}} \\ &= \frac{\cancel{\lambda^{k'} e^{-\lambda}} \cdot \frac{(n+\beta)^{n\bar{k}+\alpha}}{\Gamma(n\bar{k}+\alpha)} \lambda^{n\bar{k}+\alpha-1} e^{-(n+\beta)\lambda}}{\frac{(n+1+\beta)^{n\bar{x}+k'+\alpha}}{\Gamma(n\bar{x}+k'+\alpha)} \lambda^{n\bar{x}+k'+\alpha-1} e^{-(n+1+\beta)\lambda}} \\ &= \frac{(n+\beta)^{n\bar{k}+\alpha}}{k'! \cdot \Gamma(n\bar{k}+\alpha)} \cdot \frac{\Gamma(n\bar{x}+k'+\alpha)}{(n+1+\beta)^{n\bar{x}+k'+\alpha}} \\ &= \left(\frac{n+\beta}{n+\beta+1} \right)^{n\bar{k}+\alpha} \left(\frac{1}{n+1+\beta} \right)^{k'} \frac{\Gamma(n\bar{x}+k'+\alpha)}{k'! \cdot \Gamma(n\bar{k}+\alpha)} \\ &= \left(\frac{n+\beta}{n+\beta+1} \right)^{n\bar{k}+\alpha} \left(\frac{1}{n+1+\beta} \right)^{k'} n\bar{k} + \binom{k'+\alpha-1}{k'} \end{aligned}$$

В этом случае речь идет об отрицательном биномиальном распределении со следующими параметрами

$$P(X' | \mathbf{X}) = NB \left(n\bar{k} + \alpha, \frac{n+\beta}{n+\beta+1} \right)$$

Посмотрим на прогнозные распределения на графике (листинг 105 и рисунок 92)).

```

from scipy.stats import nbinom

a, b = 10, 2
data = np.array([5, 3, 3, 1])
n, k_bar = len(data), data.mean()

k = np.arange(1, 20)

r, p = a, b / (b + 1)
plt.plot(k, nbinom.pmf(k, r, p), 'bo', ms=8, label = 'prior predictive')
plt.vlines(k, 0, nbinom.pmf(k, r, p), colors='b', lw=5, alpha=0.5)

r_post, p_post = k_bar * n + a, (n + b)/(n + b + 1)
plt.plot(k, nbinom.pmf(k, r_post, p_post), 'go', ms=8, label = 'posterior predictive')
plt.vlines(k, 0, nbinom.pmf(k, r_post, p_post), colors='g', lw=5, alpha=0.5)

plt.title('Predictive distributions')
plt.xlabel('k')
plt.ylabel('f(k)')
plt.legend()
plt.show()

```

Листинг 105: Прогнозные распределения

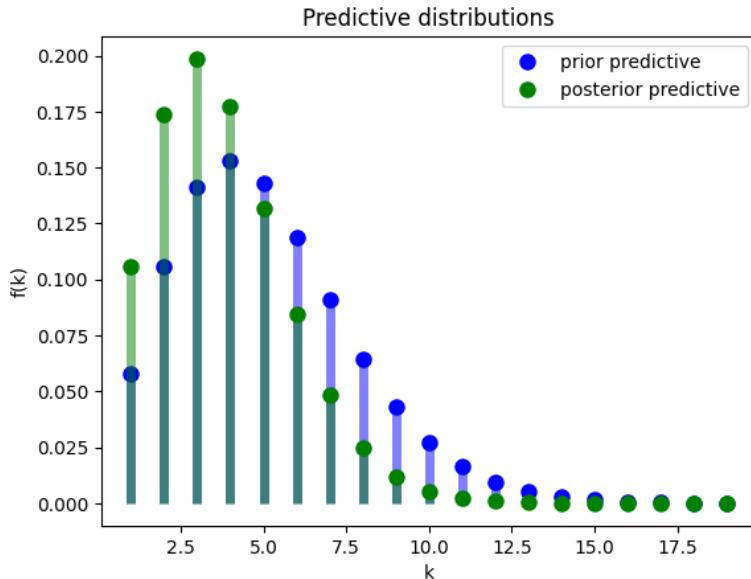


Рис. 92: Прогнозные распределения

До получения данных X разумно ожидать, что их прогнозное матожидание будет соответствовать среднему значению априорного распределения. После получения данных X ожидаемые характеристики новых данных X' должны измениться и учитывать как априорные знания о λ , так и новую информацию X (листинг 106).

```
nbinom.stats(r, p, moments = 'mv')  
Output: (5.00000000000001, 7.50000000000002)  
  
nbinom.stats(r_post, p_post, moments = 'mv')  
Output: (3.666666666666683, 4.2777777777778)
```

Листинг 106: Среднее арифметическое и дисперсия прогнозных распределений